การเปรียบเทียบเทคนิคการเรียนรู้ของเครื่องเพื่อสร้างตัวแบบการจำแนก ด้วยการปรับปรุงชุดข้อมูลอสมดุล

วริทธิ์พล แสงทองรัตนโชติ

วิทยานิพนธ์เสนอบัณฑิตวิทยาลัย มหาวิทยาลัยนเรศวร
เพื่อเป็นส่วนหนึ่งของการศึกษา หลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ
ปีการศึกษา 2565

มหาวิทยาลัยนเรศวร

การเปรียบเทียบเทคนิคการเรียนรู้ของเครื่องเพื่อสร้างตัวแบบการจำแนก ด้วยการปรับปรุงชุดข้อมูลอสมดุล

วริทธิ์พล แสงทองรัตนโชติ

วิทยานิพนธ์เสนอบัณฑิตวิทยาลัย มหาวิทยาลัยนเรศวร
เพื่อเป็นส่วนหนึ่งของการศึกษา หลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ
ปีการศึกษา 2565

วิทยานิพนธ์ เรื่อง "การเปรียบเทียบเทคนิคการเรียนรู้ของเครื่องเพื่อสร้างตัวแบบการจำแนก ด้วยการปรับปรุงชุดข้อมูลอสมดุล"

ของ วริทธิ์พล แสงทองรัตนโชติ

ได้รับการพิจารณาให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาสถิติ

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการสอบวิทยานิพนธ์

(รองศาสตราจารย์ ดร.อัชฌา อระวีพร)

..... ประธานที่ปรึกษาวิทยานิพนธ์

(รองศาสตราจารย์ ดร.อนามัย นาอุดม)

..... กรรมการที่ปรึกษาวิทยานิพนธ์

(ผู้ช่วยศาสตราจารย์ ดร.จรัสศรี รุ่งรัตนาอุบล)

..... กรรมการผู้ทรงคุณวุฒิภายใน

(ผู้ช่วยศาสตราจารย์ ดร.กัลยา บุญหล้า)

อนุมัติ

.....

(รองศาสตราจารย์ ดร.กรองกาญจน์ ชูทิพย์)

คณบดีบัณฑิตวิทยาลัย

ชื่อเรื่อง การเปรียบเทียบเทคนิคการเรียนรู้ของเครื่องเพื่อสร้างตัวแบบการจำแนก ด้วยการปรับปรุงชุดข้อมูลอสมดุล
ผู้วิจัย วริทธิ์พล แสงทองรัตนโชติ
ประธานที่ปรึกษา รองศาสตราจารย์ ดร.อนามัย นาอุดม
กรรมการที่ปรึกษา ผู้ช่วยศาสตราจารย์ ดร.จรัสศรี รุ่งรัตนาอุบล
ประเภทสารนิพนธ์ วิทยานิพนธ์ วท.ม. สถิติ, มหาวิทยาลัยนเรศวร, 2565
คำสำคัญ การวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์, ต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5, เทคนิคป่าสุ่ม, โครงข่ายประสาทเทียม, การสุ่มตัวอย่างแบบง่าย, การแบ่งกลุ่มข้อมูลแบบเคมีน, เทคนิคนาอีฟเบย์

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาประสิทธิภาพเทคนิคการจำแนกกับชุดข้อมูลที่มีจำนวนของตัวแปรอิสระเชิงคุณภาพและเชิงปริมาณแตกต่างกันทั้งหมด 3 ชุดข้อมูลได้แก่ ชุดข้อมูลสถาบันการเงินซึ่งเป็นชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพและเชิงปริมาณเท่ากัน ชุดข้อมูลสายพันธุ์ข้าวซึ่งเป็นชุดข้อมูลที่มีตัวแปรอิสระเชิงปริมาณเท่านั้นและชุดข้อมูลนักวิทยาศาสตร์ข้อมูลซึ่งเป็นชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพมากกว่าเชิงปริมาณ โดยปรับปรุงชุดข้อมูลอสมดุลให้สมดุลด้วยวิธีสุ่มลด 2 เทคนิคได้แก่ การสุ่มตัวอย่างแบบง่ายและการแบ่งกลุ่มข้อมูลแบบเคมีน แบ่งชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบด้วยหลักการ 5-Fold โดยนำชุดข้อมูลแต่ละชุดมาสร้างตัวแบบการจำแนกด้วยเทคนิคการจำแนกทั้งหมด 5 เทคนิคได้แก่ การวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์ เทคนิคนาอีฟเบย์ ต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 เทคนิคป่าสุ่มและโครงข่ายประสาทเทียม ผลจากการศึกษาพบว่า เทคนิคป่าสุ่มสามารถทำงานได้ดีภายใต้ชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณและปริมาณเท่ากัน เทคนิคการจำแนกการวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์สามารถทำงานได้ดีภายใต้ชุดข้อมูลที่มีตัวแปรอิสระเชิงปริมาณทุกตัวและโครงข่ายประสาทเทียมสามารถทำงานได้ดีภายใต้ชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพมากกว่าปริมาณและพบว่าการปรับปรุงชุดข้อมูลให้สมดุลด้วยการสุ่มตัวอย่างแบบง่ายให้ตัวแบบการจำแนกที่มีประสิทธิภาพสูงกว่าการแบ่งกลุ่มข้อมูลแบบเคมีน อีกทั้งยังพบว่าการวัดประสิทธิภาพตัวแบบการจำแนกที่สร้างจากชุดข้อมูลอสมดุล โดยใช้เพียงค่าความแม่นอย่างเดียวอาจไม่เพียงพอต่อการประเมินประสิทธิภาพ ดังนั้นควรนำค่าความเที่ยง ค่าการเรียกคืนและค่าประสิทธิภาพ มาพิจารณาประกอบการสินใจด้วย

Title COMPARISON OF MACHINE LEARNING TECHNIQUES FOR CLASSIFICATION MODEL CONSTRUCTION WITH MODIFYING IMBALANCED DATA

Author WARITPON SAENGTHONGRATTANACHOT

Advisor Associate Professor Anamai Na-udom, Ph.D.

Co-Advisor Assistant Professor Jaratsri Rungrattanaubol, Ph.D.

Academic Paper M.S. Thesis in Statistics - (Type A 2), Naresuan University, 2022

Keywords Fisher's linear discriminant analysis, Naive Bayes, Decision trees with C4.5 algorithm, Random Forest, k-mean segmentation, Simple random sampling technique, Artificial Neural Network

ABSTRACT

The purpose of this research was to study the performance of classification techniques on 3 different datasets, which are Bank dataset with an equal number of qualitative and quantitative independent variables; Data Scientist dataset with a greater number of qualitative than quantitative independent variables; and Rice Species dataset with a greater number of quantitative than qualitative. Since these datasets are imbalanced, two under sampling techniques were applied here, which are simple random sampling and k-mean clustering, to enhance the equilibrium of the data set. 5-Fold cross validation concept was applied for constructing the classification models, when designing a training dataset and test dataset. Each dataset was used to build the classification models based on 5 selected techniques including Discriminant Analysis, Naive Bayes, Decision Tree C4.5, Random Forest and Artificial Neural Network. The results indicated that Random Forest outperformed when the dataset with the same number of independent and quantitative variables. Discriminant Analysis worked well when a greater number of quantitative variables and Artificial Neural Network performed well when datasets with a greater number of qualitative variables. Moreover, the result has also shown that balancing

the dataset with simple random sampling yielded a more efficient classification model than k-mean clustering. The last notice from this study, the study confirmed that measuring the performance of imbalanced classification model with only accuracy was probably not so effective. Therefore, the precision, recall and F-measure should be considered when selecting the most appropriate classification models for making an application.

ประกาศคุณูปการ

ผู้วิจัยขอขอบพระคุณประธานที่ปรึกษาวิทยานิพนธ์ รองศาสตราจารย์ ดร.อนามัย นาอุดม เป็นอย่างสูงที่สละเวลาให้คำปรึกษาและให้คำแนะนำตลอดระยะเวลาการทำวิทยานิพนธ์ ขอขอบคุณ ผู้ช่วยศาสตราจารย์ ดร.จรัสศรี รุ่งรัตนาอุบล กรรมการที่ปรึกษาวิทยานิพนธ์ ที่ให้คำปรึกษาเกี่ยวกับการจัดการข้อมูลด้วยวิธีการทำเหมืองข้อมูล ขอขอบคุณ รองศาสตราจารย์ ดร.อัชฌา อระวีพร ประธานกรรมการสอบวิทยานิพนธ์และกรรมการผู้ทรงคุณวุฒิภายนอก ที่ให้คำแนะนำแก้ไขในส่วนที่บกพร่องของงานวิทยานิพนธ์เล่มนี้

ผู้วิจัยขอกราบขอบพระคุณคุณแม่รัตติกาล วงษ์ขาว คุณย่าสุมาลี ใจเพชร และขอบคุณเพื่อนร่วมรุ่นปริญญาโท ที่ให้การสนับสนุนงานวิทยานิพนธ์ฉบับนี้ลุล่วงไปได้ด้วยดี

ผู้วิจัยหวังเป็นอย่างยิ่งว่า งานวิจัยการเปรียบเทียบเทคนิคการเรียนรู้ของเครื่องเพื่อสร้างตัวแบบการจำแนกด้วยการปรับปรุงชุดข้อมูลอสมดุล จะเป็นประโยชน์ต่อประชาชนและบุคคลที่สนใจที่จะนำงานวิจัยนี้ไปต่อยอดเพื่อพัฒนาองค์ความรู้ใหม่ ๆ ต่อไป

วริทธิ์พล แสงทองรัตนโชติ

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญ

ในปัจจุบันไม่มีใครปฏิเสธได้ว่าข้อมูลจัดเป็นทรัพย์สินที่มีมูลค่าอย่างมากจนอาจกล่าวได้ว่า “ทรัพย์สินที่มีมูลค่ามากที่สุดในโลกไม่ใช่น้ำมันแต่เป็นข้อมูล” เนื่องจากในโลกยุคปัจจุบันทุกคนล้วนใช้เทคโนโลยีดิจิทัลในการดำเนินกิจการและชีวิตประจำวันทำให้เสมือนมีการใช้ชีวิตบนโลกออนไลน์แทบตลอดเวลา ทุกกิจกรรมบนโลกออนไลน์ย่อมก่อให้เกิดข้อมูลจำนวนมากประกอบกับความก้าวหน้าทางด้านเทคโนโลยีทำให้เกิดการสร้างข้อมูลเพิ่มขึ้นอย่างรวดเร็ว ตลอดระยะเวลาหลายปีที่ผ่านมาจำนวนข้อมูลบนโลกออนไลน์มีปริมาณเพิ่มสูงขึ้นอย่างต่อเนื่อง เมื่อข้อมูลเหล่านี้ถูกสะสมขึ้นมาเป็นจำนวนมากมายมหาศาล จนมีการนิยามคำว่า ข้อมูลขนาดใหญ่ (Big Data)

ข้อมูลขนาดใหญ่ หมายถึง ข้อมูลที่มีปริมาณมหาศาล ทำให้มีขนาดใหญ่และยากต่อการประมวลผลด้วยเทคโนโลยีแบบดั้งเดิม โดยข้อมูลขนาดใหญ่จะมีคุณสมบัติด้วยกัน 5 ประการ ได้แก่ ปริมาณ (Volume) ความหลากหลาย (Variety) ความเร็ว (Velocity) มูลค่า (Value) และความถูกต้อง (Veracity) เนื่องจากคุณสมบัติเหล่านี้จัดเป็นอุปสรรคหรือปัญหาของข้อมูลขนาดใหญ่ที่ไม่สามารถนำข้อมูลมาใช้ประโยชน์ได้ทันทีจึงต้องมีกระบวนการจัดการข้อมูลขนาดใหญ่เสียก่อน โดยกระบวนการดังกล่าวคือการทำเหมืองข้อมูล (Data Mining) ซึ่งหมายถึงการกลั่นกรองสารสนเทศที่ซ่อนอยู่ในข้อมูลขนาดใหญ่เพื่อนำสารสนเทศที่ได้มาประยุกต์ใช้เพื่อแก้ปัญหาวางแผนและช่วยในการตัดสินใจกับเหตุการณ์ที่จะเกิดขึ้นในอนาคต ซึ่งกลไกการทำเหมืองข้อมูลจะมีวิธีการทางสถิติ (Statistics) การรู้จำรูปแบบ (Pattern Recognition) และการเรียนรู้ของเครื่อง (Machine Learning) เข้ามาเกี่ยวข้อง (Burk & Miner, 2020) (อนุพงศ์ สุขประเสริฐ, 2563)

การเรียนรู้ของเครื่อง คือการให้เครื่องคอมพิวเตอร์เรียนรู้และสกัดสารสนเทศจากข้อมูลขนาดใหญ่เพื่อนำมาใช้ประโยชน์ สามารถแบ่งการเรียนรู้ของเครื่องออกเป็น 2 กลุ่มประกอบด้วย การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) และการเรียนรู้แบบมีผู้สอน (Supervised Learning)

การเรียนรู้แบบไม่มีผู้สอน คือการวิเคราะห์ข้อมูลที่ไม่มีตัวแปรตามหรือคำตอบกำกับไว้ มุ่งเน้นไปที่การหาความสัมพันธ์ระหว่างข้อมูลและการแบ่งกลุ่มเพื่อลดมิติของข้อมูล เช่น การวิเคราะห์แบ่งกลุ่ม (Cluster Analysis) ซึ่งได้รับความนิยมใช้เพื่อลดมิติของข้อมูลและการแบ่งกลุ่มข้อมูลด้วยคุณลักษณะต่าง ๆ โดยเทคนิคที่ใช้กันอย่างแพร่หลายคือ การแบ่งกลุ่มข้อมูลแบบเคมีน (k-means) เช่น การแบ่งกลุ่มข้อมูลเพื่อปรับปรุงประสิทธิภาพของระบบตรวจจับการบุกรุก (Aziz

Mohammad Nasrul & Ahmad Tohari, 2021) ซึ่งเป็นงานวิจัยที่นำวิธีการแบ่งกลุ่มข้อมูลแบบเคมีนมาประยุกต์ใช้ในงานวิจัยเพื่อลดมิติของข้อมูล

การเรียนรู้แบบมีผู้สอน คือการวิเคราะห์ข้อมูลที่มีตัวแปรตามหรือคำตอบกำกับไว้ ถ้าตัวแปรตามเป็นเชิงปริมาณจะเป็นการวิเคราะห์การถดถอย (Regression) แต่ถ้าตัวแปรตามเป็นเชิงคุณภาพจะเป็นการวิเคราะห์การจำแนก (Classification) โดยเทคนิคที่นิยมใช้อย่างแพร่หลายได้แก่ การถดถอยลอจิสติก (Logistic Regression) ต้นไม้ตัดสินใจ (Decision Tree) เทคนิคป่าสุ่ม (Random Forest) และโครงข่ายประสาทเทียม (Artificial Neural Network) เป็นต้น การจำแนกถูกนำมาประยุกต์ใช้ในงานวิจัยในหลายด้าน เช่น การศึกษาการจำแนกอัตราการเต้นหัวใจทารกในครรภ์ (Comert & Kocamaz, 2017) การศึกษาการจำแนกข้าว 2 สายพันธ์ (Cinar & Koklu, 2019) การศึกษาการจำแนกผู้ป่วยโรคความดันโลหิตสูงระดับ 2-3 (Kublanov et al., 2017) การศึกษาการจำแนกผู้ป่วยโรคหัวใจตีบ (Golpour et al., 2020) เป็นต้น

ในปัจจุบันการศึกษาเกี่ยวกับการวิเคราะห์การจำแนกเพิ่มขึ้นอย่างมากโดยงานวิจัยส่วนใหญ่มุ่งเน้นไปที่การเปรียบเทียบประสิทธิภาพตัวแบบการจำแนกมากกว่าที่จะศึกษาการเลือกใช้เทคนิคการจำแนกอย่างไรให้เหมาะสมกับชุดข้อมูล ทำให้ผู้วิจัยสนใจที่จะศึกษาประสิทธิภาพของเทคนิคการจำแนกกับชุดข้อมูลที่มีลักษณะแตกต่างกัน ซึ่งแต่ละชุดข้อมูลจะมีจำนวนตัวแปรอิสระเชิงคุณภาพและเชิงปริมาณที่แตกต่างกัน

ถึงแม้ว่าการวิเคราะห์การจำแนกจะประสบความสำเร็จอย่างมากในเชิงวิชาการและการประยุกต์ใช้ในอุตสาหกรรมต่าง ๆ แต่การวิเคราะห์การจำแนกยังพบข้อผิดพลาดและปัญหาอยู่หลายประการ หนึ่งในปัญหาที่สำคัญและไม่สามารถหลีกเลี่ยงได้คือ ความสมดุลของชุดข้อมูล โดยปัญหานี้สืบเนื่องมาจากธรรมชาติของข้อมูลมักจะมีจำนวนสมาชิกในแต่ละกลุ่มการจำแนกแตกต่างกันมากโดยกลุ่มที่มีจำนวนสมาชิกมาก (Majority Class) จะเป็นกลุ่มที่พบเจอได้ง่ายและมักจะมีจำนวนข้อมูลที่มากกว่ากลุ่มที่มีสมาชิกน้อย (Minority Class) ตัวอย่างเช่น การศึกษาการเกิดฟ้าผ่า (Pakdaman et al., 2020) โดยปรกติเหตุการณ์การเกิดฟ้าผ่าจะมีโอกาสน้อยกว่าการไม่เกิดฟ้าผ่า เนื่องจากเหตุการณ์การเกิดฟ้าผ่าเป็นเหตุการณ์ที่พบเจอได้ยากและมีจำนวนข้อมูลที่น้อยทำให้ข้อมูลคำตอบการเกิดฟ้าผ่าจัดเป็นกลุ่มที่มีจำนวนสมาชิกน้อย (Minority Class) ในขณะที่ข้อมูลคำตอบการไม่เกิดฟ้าผ่าจัดเป็นกลุ่มที่มีจำนวนสมาชิกมาก (Majority Class) ซึ่งข้อมูลลักษณะนี้คือ ข้อมูลอสมดุล (Imbalanced data) โดยจะส่งผลกระทบต่อกระบวนการทำงานของเทคนิคการจำแนกในเชิงลบ กล่าวคือเมื่อนำข้อมูลอสมดุลมาเรียนรู้การจำแนกจะส่งผลให้ตัวแบบการจำแนกมีความเอนเอียง (Bias) สามารถพยากรณ์กลุ่มที่มีสมาชิกมากได้อย่างถูกต้องและแม่นยำในขณะที่พยากรณ์กลุ่มที่มีสมาชิกน้อยผิดพลาดและไม่แม่นยำ ดังนั้นจึงมีผู้วิจัยจากหลากหลายสถาบันให้ความสำคัญและนำเสนอวิธีปรับปรุงชุดข้อมูลอสมดุลให้สมดุลก่อนนำเข้าสู่กระบวนการเรียนรู้การจำแนก โดยส่วนใหญ่มักใช้วิธีการสุ่ม

ตัวอย่างเพิ่มขึ้นของกลุ่มที่มีสมาชิกน้อยหรือสุ่มลดข้อมูลของกลุ่มที่มีสมาชิกมากก่อนนำข้อมูลเข้าสู่กระบวนการเรียนรู้การจำแนก เช่น การเปรียบเทียบประสิทธิภาพในการทำนายผลการปรับความไม่สมดุลของข้อมูลในการจำแนกด้วยเทคนิคเหมืองข้อมูล (พัชรียา ทองพูล, 2562) การเปรียบเทียบเทคนิคการสุ่มตัวอย่างเพื่อการจำแนกข้อมูลที่ไม่สมดุล (กาญจน์ ณ ศรีธะ, 2560) และการแก้ไขปัญหาข้อมูลไม่สมดุลของข้อมูลสำหรับการจำแนกผู้ป่วยโรคเบาหวาน (วิชญ์วิสฐ เกษรสิทธิ์, 2561) เป็นต้น

ปัจจุบันการศึกษาเกี่ยวกับการพัฒนากระบวนการทำงานของเทคนิคการจำแนกที่ทำงานภายใต้ข้อมูลอสมดุลมีเพิ่มมากขึ้นเรื่อย ๆ ดังนั้นการประเมินประสิทธิภาพจึงเป็นสิ่งสำคัญอย่างมาก อย่างไรก็ตามการประเมินประสิทธิภาพตัวแบบการจำแนกโดยใช้ค่าความแม่น (Accuracy) เป็นเกณฑ์เพียงเกณฑ์เดียว อาจไม่เพียงพอและไม่สามารถสะท้อนถึงข้อเท็จจริงเกี่ยวกับกระบวนการทำงานของเทคนิคการจำแนกได้อย่างครอบคลุม เพื่อให้การประเมินตัวแบบการจำแนกมีประสิทธิภาพและครอบคลุมมากขึ้น นักวิจัยมักใช้ค่าความถูกต้องอื่น ๆ มาเพื่อประกอบการตัดสินใจในการประเมินประสิทธิภาพตัวแบบการจำแนกได้แก่ ค่าความเที่ยง (Precision) ค่าการเรียกคืน (Recall) และค่าประสิทธิภาพ (F-measure) (He & Garcia, 2009)

ดังนั้นงานวิจัยนี้มีจุดประสงค์เพื่อศึกษาประสิทธิภาพของเทคนิคการจำแนกเพื่อจำแนกกลุ่มของตัวแปรตาม ภายใต้ชุดข้อมูลที่มีจำนวนของตัวแปรอิสระเชิงคุณภาพและเชิงปริมาณที่แตกต่างกัน 3 ชุดข้อมูลได้แก่ ชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพเท่ากับเชิงปริมาณ ชุดข้อมูลที่มีตัวแปรอิสระเชิงปริมาณเท่านั้นและชุดข้อมูลที่มีตัวแปรอิสระเชิงคุณภาพมากกว่าเชิงปริมาณ โดยจะประยุกต์ใช้เทคนิคการจำแนกทางด้านสถิติคือ การวิเคราะห์จำแนกกลุ่มโดยวิธีของฟิชเชอร์ (Fisher's Linear Discriminant Analysis) และใช้เทคนิคการจำแนกทางการเรียนรู้ของเครื่อง (Machine Learning) ได้แก่ เทคนิคนาอีฟเบย์ (Naïve Bayes) ต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 (Decision Tree C4.5) เทคนิคป่าสุ่ม (Random Forest) และโครงข่ายประสาทเทียม (Artificial Neural Networks) เนื่องจากเทคนิคเหล่านี้เป็นเทคนิคการจำแนกที่มีประสิทธิภาพและมีลักษณะการทำงานที่แตกต่างกันเช่น การวิเคราะห์จำแนกกลุ่มโดยวิธีของฟิชเชอร์ จะใช้ระยะห่างในการจำแนกกลุ่ม เทคนิคนาอีฟเบย์จะใช้หลักความน่าจะเป็นในการจำแนกกลุ่ม ต้นไม้ตัดสินใจจะใช้ค่าเกนสารสนเทศในการจำแนกกลุ่ม เทคนิคป่าสุ่มจะใช้การปลูกต้นไม้ตัดสินใจหลาย ๆ ต้นเพื่อช่วยในการจำแนกกลุ่ม และโครงข่ายประสาทเทียมใช้รูปแบบการคำนวณที่คล้ายระบบประสาทในการจำแนกกลุ่ม โดยในที่นี้ได้มีการเสนอแนะแนวทางการปรับปรุงชุดข้อมูลอสมดุลให้สมดุลเข้าร่วมกับการพัฒนาตัวแบบการจำแนกโดยใช้วิธีการสุ่มลด 2 เทคนิคได้แก่ การสุ่มตัวอย่างแบบง่ายและการประยุกต์ใช้การแบ่งกลุ่มข้อมูลแบบเคมีน โดยเกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพตัวแบบการจำแนกได้แก่ ค่าความแม่น ค่าความเที่ยง ค่าการเรียกคืน และค่าประสิทธิภาพ

1.2 จุดมุ่งหมายของการวิจัย

1.2.1 เพื่อศึกษาประสิทธิภาพเทคนิคการจำแนกภายใต้ชุดข้อมูล 3 ชุด ที่มีตัวแปรตามเป็นเชิงคุณภาพ โดยแต่ละชุดข้อมูลจะมีจำนวนของตัวแปรอิสระเชิงคุณภาพและเชิงปริมาณที่แตกต่างกัน

1.2.2 นำเสนอแนวทางการปรับปรุงข้อมูลอสมดุลให้สมดุลด้วยวิธีสุ่มลดประกอบด้วย การวิธีการสุ่มตัวอย่างแบบง่ายและการแบ่งกลุ่มข้อมูลแบบเคมีน

1.2.3 เพื่อเปรียบเทียบประสิทธิภาพเทคนิคการจำแนกประกอบด้วย การวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์ เทคนิคนาอีฟเบย์ ต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 เทคนิคป่าสุ่ม และโครงข่ายประสาทเทียม

1.3 ขอบเขตการวิจัย

1.3.1 ขอบเขตด้านข้อมูล

งานวิจัยนี้มีจุดประสงค์เพื่อศึกษาประสิทธิภาพเทคนิคการจำแนกภายใต้ชุดข้อมูลที่มีจำนวตัวตัวแปรอิสระที่แตกต่างกัน โดยแต่ละชุดข้อมูลจะมีจำนวนตัวแปรอิสระเชิงคุณภาพและเชิงปริมาณที่แตกต่างกัน ผู้วิจัยคัดเลือกชุดข้อมูลทั้ง 3 ชุด ซึ่งมีแหล่งที่มาจาก UCI Machine Learning Repository โดยแต่ละชุดข้อมูลมีรายละเอียดดังต่อไปนี้

1. ชุดข้อมูลที่ 1 คือ ข้อมูลสถาบันการเงินมีตัวแปรอิสระ (Independent Variable) จำนวน 20 ตัว ประกอบด้วยตัวแปรอิสระเชิงคุณภาพ 10 ตัวและตัวแปรอิสระเชิงปริมาณ 10 ตัว โดยมีตัวแปรตาม (Dependent Variable) เชิงคุณภาพ 1 ตัวซึ่งแบ่งออกเป็น 2 กลุ่ม

2. ชุดข้อมูลที่ 2 คือ ข้อมูลสายพันธ์ข้าวมีตัวแปรอิสระจำนวน 7 ตัวประกอบด้วยตัวแปรอิสระเชิงคุณภาพ 0 ตัวและตัวแปรอิสระเชิงปริมาณ 7 ตัว โดยมีตัวแปรตามเชิงคุณภาพ 1 ตัวซึ่งแบ่งออกเป็น 2 กลุ่ม

3. ชุดข้อมูลที่ 3 คือ ข้อมูลนักวิทยาศาสตร์ข้อมูลมีตัวแปรอิสระจำนวน 10 ตัวประกอบด้วยตัวแปรเชิงอิสระเชิงคุณภาพ 9 ตัวและตัวแปรอิสระเชิงปริมาณ 1 ตัว โดยมีตัวแปรตามเชิงคุณภาพ 1 ตัวซึ่งแบ่งออกเป็น 2 กลุ่ม

1.3.2 ขอบเขตด้านเทคนิคการจำแนก

จากการทบทวนวรรณกรรมผู้วิจัยได้ทำการคัดเลือกเทคนิคการจำแนกที่ได้รับความนิยมและมีประสิทธิภาพมาประยุกต์ใช้ในงานวิจัยนี้ ได้แก่

1. การวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์
2. เทคนิคนาอีฟเบย์
3. ต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5
4. เทคนิคป่าสุ่ม
5. โครงข่ายประสาทเทียม

1.3.3 ขอบเขตด้านการปรับปรุงชุดข้อมูลอสมดุลให้สมดุล

งานวิจัยนี้จะปรับปรุงชุดข้อมูลอสมดุลให้สมดุลโดยใช้วิธีสุ่มลด (Under-Sampling) โดยแบ่งออกเป็น 2 เทคนิคได้แก่ การสุ่มตัวอย่างแบบง่าย (Random Under-Sampling) และการแบ่งกลุ่มข้อมูลแบบเคมีน (k-means)

1.3.4 ขอบเขตด้านการออกแบบชุดข้อมูลเรียนรู้ (Training set) และชุดข้อมูลทดสอบ (Test set)

งานวิจัยนี้จะใช้หลักการ 5-Fold Cross Validation ซึ่งเป็นการแบ่งชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบด้วยอัตราส่วน 80 ต่อ 20 และสร้างตัวแบบการจำแนกซ้ำ 5 ครั้ง

1.3.5 ขอบเขตเครื่องมือที่ใช้วัดประสิทธิภาพตัวแบบการจำแนก

1. ค่าความแม่น (Accuracy)
2. ค่าความเที่ยง (Precision)
3. ค่าการเรียกคืน (Recall)
4. ค่าประสิทธิภาพ (F-measure)

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. ทราบแนวทางการเลือกใช้เทคนิคการจำแนกให้เหมาะสมภายใต้ชุดข้อมูลลักษณะต่าง ๆ
2. ทราบแนวทางการปรับปรุงชุดข้อมูลอสมดุลให้สมดุล

1.5 นิยามศัพท์เฉพาะ

ในการนำเสนอผลการวิจัยสำหรับงานวิจัยนี้ เพื่อให้เกิดความเข้าใจตรงกันในการแปลความหมาย ดังนั้นผู้วิจัยจึงใช้สัญลักษณ์และอักษรย่อดังนี้

TP	แทน	ค่าความถูกต้องเชิงบวก (True Positive)
TN	แทน	ค่าความถูกต้องเชิงลบ (True Negative)
FP	แทน	ค่าความผิดพลาดเชิงบวก (False Positive)
FN	แทน	ค่าความผิดพลาดเชิงลบ (False Negative)
Acc	แทน	ค่าความแม่น (Accuracy)
Pre	แทน	ค่าความเที่ยง (Precision)
Re	แทน	ค่าการเรียกคืน (Recall)
F	แทน	ค่าประสิทธิภาพ (F-measure)
LDA	แทน	ตัวแบบการจำแนกของเทคนิคการวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์
NB	แทน	ตัวแบบการจำแนกของเทคนิคนาอีฟเบย์

DT	แทน	ตัวแบบการจำแนกของเทคนิคต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5
RF	แทน	ตัวแบบการจำแนกของเทคนิคป่าสุ่ม
ANN	แทน	ตัวแบบการจำแนกของเทคนิคโครงข่ายประสาทเทียม
Rank	แทน	การเรียงลำดับตัวแบบการจำแนกที่มีประสิทธิภาพมากที่สุดไปหาน้อยที่สุด
Average	แทน	ค่าเฉลี่ย

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

งานวิจัยนี้มีจุดประสงค์เพื่อเปรียบเทียบประสิทธิภาพตัวแบบการจำแนกทางด้านสถิติและทางด้านการเรียนรู้ของเครื่องและนำเสนอแนวทางการปรับปรุงชุดข้อมูลอสมดุลให้สมดุลโดยมีเอกสารและงานวิจัยที่เกี่ยวข้องดังนี้

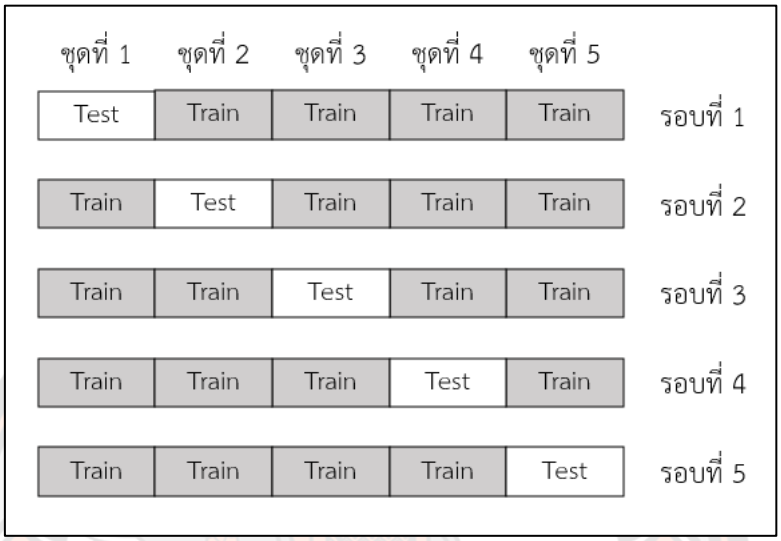
2.1 การพัฒนาตัวแบบการจำแนก

การพัฒนาตัวแบบการจำแนก ในขั้นตอนแรกจะต้องนำชุดข้อมูล (Dataset) มาแบ่งออกเป็นชุดข้อมูล 2 ชุดได้แก่ ชุดข้อมูลเรียนรู้ (Training set) และชุดข้อมูลทดสอบ (Test set) และนำชุดข้อมูลเรียนรู้ไปสร้างตัวแบบการจำแนกและนำตัวแบบการจำแนกที่ได้ไปทดสอบกับชุดข้อมูลทดสอบ ดังนั้นประสิทธิภาพของตัวแบบการจำแนกจะวัดจากความถูกต้องในการพยากรณ์การจำแนกภายใต้ชุดข้อมูลทดสอบเท่านั้น

โดยทั่วไปการแบ่งชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบสามารถทำได้โดยการสุ่มอิสระตามสัดส่วนต่าง ๆ ตามที่ผู้ศึกษาต้องการ เช่น แบ่งข้อมูลออกเป็นชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ 80:20 และ 70:30 เป็นต้น โดยชุดข้อมูลเรียนรู้จะมีจำนวนมากกว่าชุดข้อมูลทดสอบเสมอ เนื่องจากการแบ่งชุดข้อมูลด้วยวิธีดังกล่าวก็ยังมีการตั้งคำถามถึงประสิทธิภาพในชุดข้อมูลเรียนรู้ว่าเป็นชุดข้อมูลเรียนรู้ที่ครอบคลุมคุณลักษณะที่สำคัญของชุดข้อมูลหรือไม่ ดังนั้นถ้าไม่ใช่ชุดข้อมูลเรียนรู้ที่ดีพอจะส่งผลต่อประสิทธิภาพของตัวแบบการจำแนกได้

เพื่อให้ได้การแบ่งชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบที่มีความรอบคอบและมีประสิทธิภาพโดยไม่ยึดติดกับชุดข้อมูลใดชุดข้อมูลหนึ่ง ผู้พัฒนาส่วนใหญ่จะทำการสุ่มแบ่งชุดข้อมูลเป็นหลายชุดและสร้างตัวแบบการจำแนกหลายตัวแบบโดยหลักการทำงานในลักษณะนี้คือหลักการทำงาน k-Fold (k-Fold Cross-Validation) โดยกลไกการทำงานจะมีการแบ่งชุดข้อมูลเป็น k กลุ่มและนำข้อมูลจำนวน k-1 กลุ่มมาเป็นชุดข้อมูลเรียนรู้ โดยหนึ่งกลุ่มที่เหลือจะใช้เป็นชุดข้อมูลทดสอบโดยจะมีการวนสับเปลี่ยนกลุ่มเหล่านี้จะทำให้ในกระบวนการมีการสร้างตัวแบบการจำแนกและนำตัวแบบไปทดสอบทั้งหมด k รอบ ดังภาพที่ 1 ซึ่งเป็นการทำงานแบบ 5-Fold โดยมีรายละเอียดดังนี้ จากชุดข้อมูล 100% จะถูกแบ่งออกเป็น 5 กลุ่ม กลุ่มละ 20% จากนั้นในรอบที่ 1 จะให้ชุดข้อมูลกลุ่มที่ 1 เป็นชุดข้อมูลทดสอบ และชุดข้อมูลกลุ่มที่ 2-5 เป็นชุดข้อมูลเรียนรู้ (คล้ายการแบ่งชุดข้อมูล 80:20) จากนั้นสับเปลี่ยนชุดข้อมูลจนครบ 5 รอบ นำค่าความถูกต้องของแต่ละรอบมาคำนวณหาค่าเฉลี่ยเพื่อแสดงประสิทธิภาพของตัวแบบการจำแนก จะเห็นได้ว่าวิธีการแบ่งชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบแบบ k-Fold มีความละเอียดและรอบคอบกว่า

วิธีก่อนหน้า เนื่องจากมีการพัฒนาตัวแบบการจำแนกหลายรอบ ซึ่งต่างจากการแบ่งกลุ่มข้อมูลด้วยวิธีก่อนหน้าที่มีการเรียนรู้เพียงแค่รอบเดียว โดยจะวัดประสิทธิภาพการจำแนกกับชุดข้อมูลทดสอบ



ภาพที่ 1 โครงสร้างการทำงานของ 5-Fold Cross Validation

การวัดประสิทธิภาพของตัวแบบการจำแนก

เครื่องมือที่ใช้วัดประสิทธิภาพตัวแบบการจำแนกในที่นี้มีด้วยกัน 4 ตัว ดังที่กำหนดไว้ก่อนหน้าผ่านผลลัพธ์การทำงานด้วยเมทริกซ์ความสับสน (Confusion Matrix)

เมทริกซ์ความสับสน (Confusion Matrix)

เมทริกซ์ความสับสน คือตารางแสดงผลลัพธ์การทำนายของตัวแบบการจำแนก โดยนำผลลัพธ์ของตัวแบบการจำแนกเปรียบเทียบกับผลลัพธ์จริง โดยมีส่วนประกอบได้แก่ ค่า True Positive (TP) ค่า True Negative (TN) ค่า False Positive (FP) และค่า False Negative (FN) แสดงดังตารางที่ 1

ตารางที่ 1 เมทริกซ์ความสับสน (Confusion Matrix)

ค่าจริง (Actual Class)	ค่าทำนาย (Predicted Class)	
	Yes	No
Yes	TP	FN
No	FP	TN

อธิบายรายละเอียดจากตารางที่ 1 ได้ดังนี้

TP (True Positive) คือ จำนวนครั้งที่การจำแนกข้อมูลซึ่งมีค่าจริงอยู่ใน Class Yes และมีการทำนายว่าอยู่ใน Class Yes (ทำนายถูกต้อง)

FN (False Negative) คือ จำนวนครั้งที่การจำแนกข้อมูลซึ่งมีค่าจริงอยู่ใน Class Yes แต่มีการทำนายว่าอยู่ใน Class No (ทำนายผิด คำตอบจริงเป็น Yes แต่ทำนายว่าเป็น No)

FP (False Positive) คือ จำนวนครั้งที่การจำแนกข้อมูลซึ่งมีค่าจริงอยู่ใน Class No แต่มีการทำนายว่าอยู่ใน Class Yes (ทำนายผิด คำตอบจริงเป็น No แต่ทำนายเป็น Yes)

TN (True Negative) คือ จำนวนครั้งที่การจำแนกข้อมูลซึ่งมีค่าจริงอยู่ใน Class No และมีการทำนายว่าอยู่ใน Class No (ทำนายถูกต้อง)

นำค่าที่ได้จากเมทริกซ์ความสับสนมาคำนวณเพื่อหาค่าความแม่น ค่าความเที่ยง ค่าการเรียกคืน ค่าวัดประสิทธิภาพ ดังต่อไปนี้

ค่าความแม่น (Accuracy) คือค่าที่แสดงจำนวนครั้งที่พยากรณ์ถูกต้องต่อจำนวนข้อมูลทั้งหมด หรือกล่าวอีกนัยหนึ่งคือ ค่าร้อยละความถูกต้องของการพยากรณ์ต่อจำนวนครั้งการพยากรณ์ทั้งหมด โดยสามารถคำนวณได้สมการที่ 1

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \tag{1}$$

ค่าความเที่ยง (Precision) คือ ค่าที่อธิบายถึงความถูกต้องของกลุ่มข้อมูลที่กำลังพิจารณาเมื่อเทียบกับผลลัพธ์ของการทำนาย สามารถคำนวณได้สมการที่ 2 และสมการที่ 3

$$Precision(Yes) = \frac{TP}{(TP + FP)} \tag{2}$$

$$Precision(No) = \frac{TN}{(TN + FN)} \tag{3}$$

ค่าการเรียกคืน (Recall) คือ ค่าที่อธิบายถึงความถูกต้องของผลการทำนายของกลุ่มข้อมูลที่กำลังพิจารณาอยู่เมื่อเทียบกับผลของความเป็นจริง สามารถคำนวณได้สมการที่ 4 และสมการที่ 5

$$Recall(Yes) = \frac{TP}{(TP + FN)} \tag{4}$$

$$Recall(No) = \frac{TN}{(TN + FP)} \tag{5}$$

ค่าประสิทธิภาพ (F-measure) คือ ค่าเฉลี่ยของค่ากลางของผลจากการหารจำนวนข้อมูลทั้งหมด สามารถคำนวณได้สมการที่ 6 และ 7

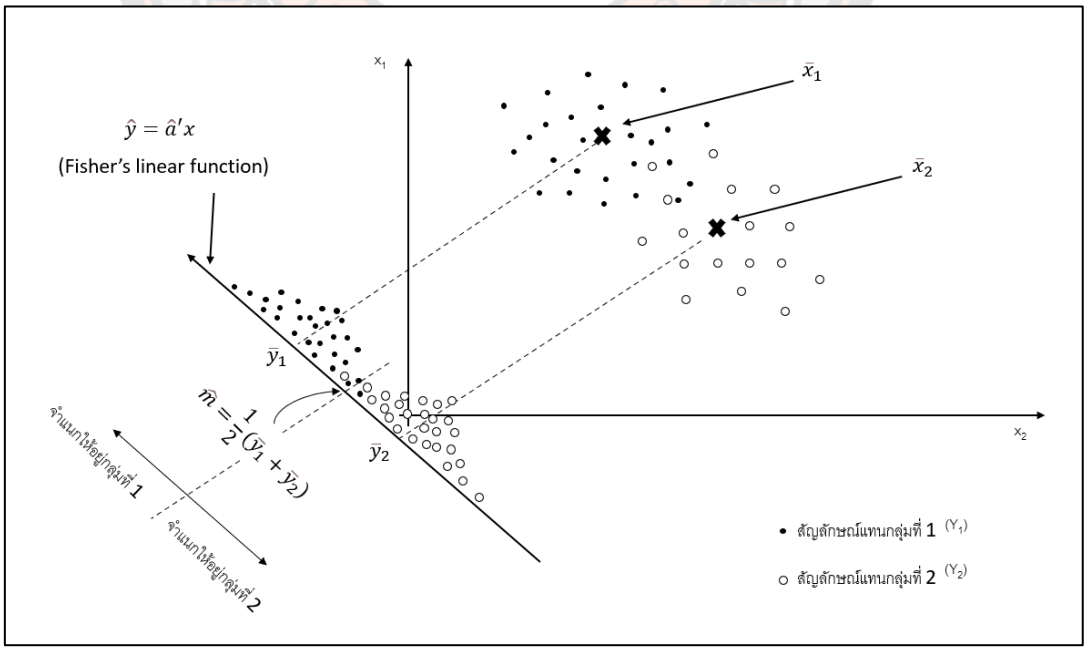
$$F-measure(Yes)=\frac{2\times Precision(Yes)\times Recall(Yes)}{Precision(Yes)+Recall(Yes)} \quad (6)$$

$$F-measure(No)=\frac{2\times Precision(No)\times Recall(No)}{Precision(No)+Recall(No)} \quad (7)$$

2.2 เทคนิคการจำแนกที่ใช้ในงานวิจัย

2.2.1 การวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์ (Fisher's Linear Discriminant Analysis)

การวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์เป็นหนึ่งในเทคนิคการจำแนกกลุ่มข้อมูลทางด้านสถิติ โดยพิจารณาการจำแนกกลุ่มด้วยระยะห่างระหว่างข้อมูลกับค่ากลางของกลุ่มผ่านสมการเชิงเส้นโดยวิธีของฟิชเชอร์ (Fisher's linear function) ดังภาพที่ 2 ซึ่งสมการจะอยู่ในรูปแบบสมการเชิงเส้น และต้องมีอัตราส่วนความผันแปรระหว่างกลุ่ม (Sum Square Between group) กับความผันแปรภายในกลุ่ม (Sum Square Within group) มีค่ามากที่สุด โดยตั้งต้นการวิเคราะห์จำแนกกลุ่มเชิงเส้นจะมีเงื่อนไขหรือข้อตกลง (Assumption) อยู่ 2 ข้อประกอบด้วย ตัวแปรอิสระมีการแจกแจงแบบปรกติหลายตัวแปร (Normality of Independent Variables) และเมทริกซ์ความแปรปรวนร่วมของตัวแปรอิสระต้องเท่ากัน (Equal Dispersion Matrices) แต่การวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์จะไม่พิจารณาว่าประชากรจะมีการแจกแจงแบบปรกติหรือไม่และสมมติโดยปริยายว่าเมทริกซ์ความแปรปรวนของประชากรระหว่างกลุ่มเท่ากัน (Johnson, 2014)



ภาพที่ 2 การวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์

อ้างอิง : ดัดแปลงจากหนังสือ Applied Multivariate Statistical Analysis (Johnson & Wichern, 2014)

จากภาพที่ 2 แสดงการวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์ พิจารณาจำแนกกลุ่มโดยใช้เส้นตรง $\hat{m}$ เป็นเกณฑ์กล่าวคือนำเมื่อชุดข้อมูลใหม่คำนวณผ่านสมการเชิงเส้นโดยวิธีของฟิชเชอร์หากค่าที่ได้มีค่ามากกว่าเส้นตรง $\hat{m}$ จะจำแนกให้อยู่กลุ่มที่ 1 ในทางตรงกันข้ามหากมีค่าน้อยกว่าเส้นตรง $\hat{m}$ จะจำแนกให้อยู่กลุ่มที่ 2 ซึ่งการคำนวณหาเส้นตรง $\hat{m}$ จะคำนวณผ่านสมการเชิงเส้นโดยวิธีของฟิชเชอร์ โดยสมการเชิงเส้นโดยวิธีของฟิชเชอร์สามารถคำนวณได้ดังสมการที่ 8

$$\begin{aligned} \hat{y} &= \hat{b}'x \\ &= \left[\overline{x}_1 - \overline{x}_2\right] S_{pooled}^{-1} x \end{aligned} \tag{8}$$

เมื่อ

$$S_{pooled} = \left[\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}\right] S_1 + \left[\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}\right] S_2 \tag{9}$$

$$\underset{(p\times p)}{\overline{x}_1} = \frac{1}{n_1}\sum_{j=1}^{n_1} x_{1j} \tag{10}$$

$$\underset{(p\times p)}{\overline{x}_2} = \frac{1}{n_2}\sum_{j=1}^{n_2} x_{1j} \tag{11}$$

$$\underset{(p\times p)}{S_1} = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}\left(x_{1j} - \overline{x}_1\right)\left(x_{1j} - \overline{x}_1\right)' \tag{12}$$

$$\underset{(p\times p)}{S_2} = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}\left(x_{2j} - \overline{x}_2\right)\left(x_{2j} - \overline{x}_2\right)' \tag{13}$$

เมื่อ

$\hat{y}$ คือ สมการเชิงเส้นโดยวิธีของฟิชเชอร์

x คือ เวกเตอร์ของตัวแปรอิสระ p ตัว

$\hat{b}'$ คือ เวกเตอร์สัมประสิทธิ์จำแนกประเภท

$\overline{x}_j$ คือ เวกเตอร์ค่าเฉลี่ยตัวแปรอิสระในกลุ่มที่ j

S_j คือ เมทริกซ์ของค่าความแปรปรวนร่วมกลุ่มที่ j

S_{pooled} คือ เมทริกซ์ค่าแปรปรวนร่วมตัวอย่างรวมกัน

n_j คือ จำนวนหน่วยตัวอย่างของกลุ่มที่ j

ในหัวข้อถัดไปผู้วิจัยจะแสดงตัวอย่างการคำนวณการจำแนกด้วยเทคนิคการวิเคราะห์กลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์ ด้วยชุดข้อมูลตัวอย่างที่เป็นข้อมูลความเสี่ยงในการให้เครดิต ซึ่งเป็นชุดข้อมูลที่ผู้วิจัยสร้างขึ้นมา ดังตารางที่ 2

ตารางที่ 2 ข้อมูลการประเมินความเสี่ยงเครดิตของลูกค้าชุดที่ 1

Customer	Assets (x_1)	Income (x_2)	Credit Risk (Y)
1	High	50	Good
2	High	100	Good
3	Low	25	Good
4	High	75	Good
5	Low	75	Good
6	Low	25	Bad
7	Low	50	Bad
8	Low	25	Bad

จากตารางที่ 2 เนื่องจากข้อมูลชุดนี้มีตัวแปรอิสระ Assets ซึ่งเป็นตัวแปรเชิงคุณภาพ ทำให้ก่อนการคำนวณจะต้องแปลงค่าของตัวแปรนี้ให้เป็นตัวแปรดัมมี่เสียก่อน ทำให้ได้ค่าดังตารางต่อไปนี้

Customer	Assets = High (x_1)	Income (x_2)	Credit Risk (Y)
1	1	50	Good
2	1	100	Good
3	0	25	Good
4	1	75	Good
5	0	75	Good
6	0	25	Bad
7	0	50	Bad
8	0	25	Bad

ในที่นี้จะกำหนดให้ Credit Risk = Good คือกลุ่ม Y_1 และ Credit Risk = Bad คือกลุ่ม Y_2

ขั้นตอนที่ 1 คำนวณค่าเฉลี่ยของตัวแปรแต่ละกลุ่มและเมทริกซ์ค่าแปรปรวนร่วมตัวอย่างรวมกัน ดังนี้

$$\bar{x}_1 = \begin{bmatrix} \left(\frac{1+1+1+0+0}{5}\right) \\ \left(\frac{75+50+100+25+75}{5}\right) \end{bmatrix} = \begin{bmatrix} 0.6 \\ 65 \end{bmatrix}, \quad \bar{x}_2 = \begin{bmatrix} \left(\frac{0+0+0}{3}\right) \\ \left(\frac{25+50+25}{3}\right) \end{bmatrix} = \begin{bmatrix} 0 \\ 33.33 \end{bmatrix}$$

และ $$S_{pooled}^{-1} = \begin{bmatrix} 0.15 & 107.615 \\ 107.615 & 406.25 \end{bmatrix}$$

ขั้นตอนที่ 2 คำนวณหาสมการเชิงเส้นโดยวิธีของฟิชเชอร์ได้ดังนี้

$$\begin{aligned} \hat{y} &= \hat{b}'x \\ &= [\bar{x}_1 - \bar{x}_2] S_{pooled}^{-1} x \\ &= \begin{bmatrix} 0.6 & 31.67 \end{bmatrix} \begin{bmatrix} 0.15 & 107.615 \\ 107.615 & 406.25 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= 0.264x_1 + 0.0054x_2 \end{aligned}$$

ขั้นตอนที่ 3 คำนวณหาค่ากลางของกลุ่ม Y_1 และค่ากลางของกลุ่ม Y_2 ดังนี้

$$\bar{y}_1 = \hat{b}'\bar{x}_1 = \begin{bmatrix} 0.264 & 0.0054 \end{bmatrix} \begin{bmatrix} 0.6 \\ 65 \end{bmatrix} = 0.5094$$

$$\bar{y}_2 = \hat{b}'\bar{x}_2 = \begin{bmatrix} 0.264 & 0.0054 \end{bmatrix} \begin{bmatrix} 0 \\ 33.33 \end{bmatrix} = 0.18$$

ขั้นตอนที่ 4 คำนวณหาค่า $\hat{m}$ ได้ดังนี้

$$\hat{m} = \frac{1}{2}\left(\bar{y}_1 + \bar{y}_2\right) = \frac{1}{2}\left(0.5094 + 0.18\right) = 0.3447$$

ดังนั้นใช้ค่า $\hat{m}$ เป็นเกณฑ์ในการจำแนกกลุ่ม โดยมีเกณฑ์ดังต่อไปนี้

จะจำแนกให้อยู่กลุ่มที่ 1 หาก $\hat{y} = \hat{b}'x_0 \geq \hat{m} = 0.3447$

จะจำแนกให้อยู่กลุ่มที่ 2 หาก $\hat{y} = \hat{b}'x_0 < \hat{m} = 0.3447$

ในหัวข้อถัดไปผู้วิจัยจะยกตัวอย่างการนำเกณฑ์ข้างต้นมาใช้ในการพยากรณ์จำแนกกลุ่ม โดยกำหนดชุดข้อมูลใหม่ที่ต้องการพยากรณ์คือ x_0 = (Assets = 0, Income = 65) จงพยากรณ์ว่าข้อมูลใหม่ดังกล่าวควรจำแนกให้ Credit Risk อยู่กลุ่มใด

นำชุดข้อมูลใหม่มาคำนวณผ่านสมการเชิงเส้นโดยวิธีของฟิชเชอร์ได้ดังนี้

$$\begin{aligned} \hat{y} &= 0.264x_1 + 0.0054x_2 \\ &= (0.264)\,(0) + (0.0054)\,(65) = 0.351 \end{aligned}$$

เนื่องจากค่าที่คำนวณได้มีค่าเท่ากับ $\hat{y}$ = 0.351 ซึ่งมีค่ามากกว่า 0.3447 ดังนั้นจะจำแนกให้ข้อมูลดังกล่าวอยู่กลุ่ม Y_1 นั่นคือจำแนกให้ Credit Risk อยู่กลุ่ม Good

2.2.2 เทคนิคนาอีฟเบย์ (Naïve Bayes)

เทคนิคนาอีฟเบย์เป็นหนึ่งในเทคนิคทางด้านการเรียนรู้ของเครื่องซึ่งอาศัยหลักความน่าจะเป็นบนทฤษฎีบทของเบย์โดยจะกำหนดให้แต่ละเหตุการณ์เป็นผลแบ่งกั้น (partition) ของปริภูมิตัวอย่าง หรือกล่าวอีกนัยหนึ่งคือแต่ละเหตุการณ์จะไม่เกิดร่วมกันทำให้การเกิดของเหตุการณ์ต่าง ๆ ที่ใช้ในการจำแนกกลุ่มนั้นเป็นอิสระต่อกันโดยมีสมการความน่าจะเป็นแบบมีเงื่อนไขดังสมการที่ 14

$$P\left(Y\middle|X\right)=\frac{P\left(X\middle|Y\right)P(Y)}{P(X)} \tag{14}$$

จากสมการที่ 14 กำหนดให้ *P(Y)* คือความน่าจะเป็นที่จะเกิดเหตุการณ์ *Y* และ *P(Y|X)* คือความน่าจะเป็นที่จะเกิดเหตุการณ์ *Y* เมื่อเกิดเหตุการณ์ *Y* ก่อนหน้าแล้ว

ตัวจำแนกแบบเบย์จะประยุกต์นำทฤษฎีของเบย์มาช่วยในการจำแนกกลุ่มแต่ในการวิเคราะห์ร่วมกับข้อมูลขนาดใหญ่ซึ่งในชุดข้อมูลจะมีตัวแปรอิสระจำนวนมากดังนั้นจะได้สมการความน่าจะเป็นแบบมีเงื่อนไขดังสมการที่ 15

$$P\left(Y_j\mid x_1,x_2,...,x_p\right)=\frac{P\left(x_1,x_2,...,x_p\mid Y\right)P(Y_j)}{P\left(x_1,x_2,...,x_p\right)} \tag{15}$$

โดย Y_i คือ ตัวแปรตามกลุ่มที่ j เมื่อ j = 1, 2, ..., k

x_i คือ ตัวแปรอิสระที่ i เมื่อ i = 1, 2, ..., p

โดยการวิเคราะห์จำแนกกลุ่มจะใช้สมการความน่าจะเป็นดังสมการที่ 16

$$P(Y_j\left|x_1,x_2,...,x_p\right.)P(Y_j)=P(x_1\left|Y_j\right.)P(x_2\left|Y_j\right.)...P(x_p\left|Y_j\right.)P(Y_j) \tag{16}$$

การพยากรณ์การจำแนกกลุ่มจะใช้สมการที่ 16 โดยพิจารณาทีละกลุ่ม กลุ่มใดมีค่าความน่าจะเป็นสูงที่สุดจะพิจารณาให้อยู่กลุ่มนั้นหรือกล่าวอีกนัยหนึ่งคือการหาค่า Maximize $P(Y_j\left|x_1,x_2,...,x_p\right.)\,P(Y_j)$

ในหัวข้อต่อไปผู้วิจัยจะแสดงตัวอย่างการคำนวณการจำแนกโดยใช้เทคนิคนาอีฟเบย์ โดยใช้ข้อมูลความเสี่ยงในการให้เครดิต ซึ่งเป็นชุดข้อมูลที่ผู้วิจัยสร้างขึ้นมา ดังตารางที่ 3

ตารางที่ 3 ข้อมูลการประเมินความเสี่ยงเครดิตของลูกค้าชุดที่ 2

Customer	Saving (x_1)	Assets (x_2)	Income (x_3)	Credit Risk
1	Medium	High	75	Good
2	Low	Low	50	Bad
3	High	Low	25	Bad
4	Medium	High	50	Good
5	Low	High	100	Good
6	High	Low	25	Good
7	Low	Low	25	Bad
8	Medium	Low	75	Good

กำหนดให้ชุดข้อมูลใหม่ที่ต้องการพยากรณ์การจำแนกกลุ่มคือ x=(Saving=High, Assets=Low, Income=65) จงพยากรณ์ว่าชุดข้อมูลใหม่ซึ่งมีคุณลักษณะดังกล่าวควรจำแนกให้ Credit Risk อยู่กลุ่มใด

ในที่นี้จะกำหนดให้ Credit Risk = Good คือกลุ่ม Y_1 และCredit Risk = Bad คือกลุ่ม Y_2

ขั้นตอนที่ 1 คำนวณความน่าจะเป็นแต่ละกลุ่มของตัวแปรตาม (Y_j) ดังนี้

$$P(Y_1) = \frac{5}{8},\ P(Y_2) = \frac{3}{8}$$

ขั้นตอนที่ 2 คำนวณความน่าจะเป็นตัวแปรอิสระโดยพิจารณาแยกแต่ละกลุ่ม ดังนี้

คำนวณความน่าจะเป็นของตัวแปร Saving (x_1) = High ดังนี้

$$P(x_1 = High \mid Y_1) = \frac{1}{5},\ P(x_1 = High \mid Y_2) = \frac{1}{3}$$

คำนวณความน่าจะเป็นของตัวแปร Assets ($x2$) = Low ดังนี้

$$P(x_2 = Low \mid Y_1) = \frac{2}{5},\ P(x_2 = Low \mid Y_2) = \frac{3}{3}$$

การจำแนกกลุ่มข้อมูลด้วยเทคนิคนาอีฟเบย์ในกรณีที่ชุดข้อมูลมีตัวแปรอิสระเชิงปริมาณ ในการคำนวณความน่าจะเป็นจะอาศัยแนวคิดของการแจกแจงแบบปรกติ (Normal Distribution) ดังต่อไปนี้

$$P(Y_j \mid x_i) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{1}{2\sigma_{ij}^2}\left(x_i - \mu_{ij}\right)^2}$$

โดยที่ μ_{ij} คือ ค่าเฉลี่ยของข้อมูลตัวแปรอิสระที่ i และตัวแปรตามกลุ่มที่ j

σ_{ij} คือ ส่วนเบี่ยงเบนมาตรฐานตัวแปรอิสระที่ i และตัวแปรตามกลุ่มที่ j

คำนวณความน่าจะเป็นของตัวแปร Income (x_3) = 65 ในกรณีที่ Credit Risk เป็นกลุ่ม Y_1 ดังนี้

โดยที่ x_3 = 65 (ค่าของตัวแปร Income)

μ_{31} = 65 (ค่าเฉลี่ยของตัวแปร Income และ Credit Risk เป็นกลุ่ม Y_1)

σ_{31} = 25.495 (ส่วนเบี่ยงเบนมาตรฐานของตัวแปร Income และ Credit Risk เป็นกลุ่ม Y_1)

จะได้

$$P(x_3 = 65 \mid Y_1) = \frac{1}{\sqrt{2\pi}(25.495)} e^{-\frac{1}{2(25.495)^2}(65-65)^2}$$

$$= 0.0157$$

คำนวณความน่าจะเป็นของตัวแปร Income (x_3) = 65 ในกรณีที่ Credit Risk เป็นกลุ่ม Y_2 ดังนี้

โดยที่ x_3 = 65 (ค่าของตัวแปร Income)

μ_{32} = 33.33 (ค่าเฉลี่ยของตัวแปร Income และ Credit Risk เป็นกลุ่ม Y_2)

σ_{32} = 11.785 (ส่วนเบี่ยงเบนมาตรฐานของตัวแปร Income และ Credit Risk เป็นกลุ่ม Y_2)

จะได้

$$P(x_3 = 65 \mid Y_2) = \frac{1}{\sqrt{2\pi}(11.785)} e^{-\frac{1}{2(11.785)^2}(65-33.33)^2}$$

$$= 0.000914$$

ขั้นตอนที่ 3 คำนวณหาความน่าจะเป็นเพื่อพยากรณ์การจำแนกกลุ่ม

คำนวณความน่าจะเป็นเพื่อจำแนกให้ชุดข้อมูลดังกล่าวอยู่กลุ่ม Y_1 ดังนี้

$$= P(Y_1 \mid x_1 = High, x_2 = Low, x_3 = 65)P(Y_1)$$

$$= P(x_1 = High \mid Y_1)P(x_2 = Low \mid Y_1)P(x_3 = 65 \mid Y_1)P(Y_1)$$

$$= \left(\frac{1}{5}\right)\left(\frac{2}{5}\right)(0.0157)\left(\frac{5}{8}\right)$$

$$= 0.00785$$

คำนวณความน่าจะเป็นเพื่อจำแนกให้ข้อมูลดังกล่าวอยู่กลุ่ม Y_2 ดังนี้

$$= P(Y_2 \mid x_1 = High, x_2 = Low, x_3 = 65)P(Y_2)$$

$$= P(x_1 = High \mid Y_2)P(x_2 = Low \mid Y_2)P(x_3 = 65 \mid Y_2)P(Y_2)$$

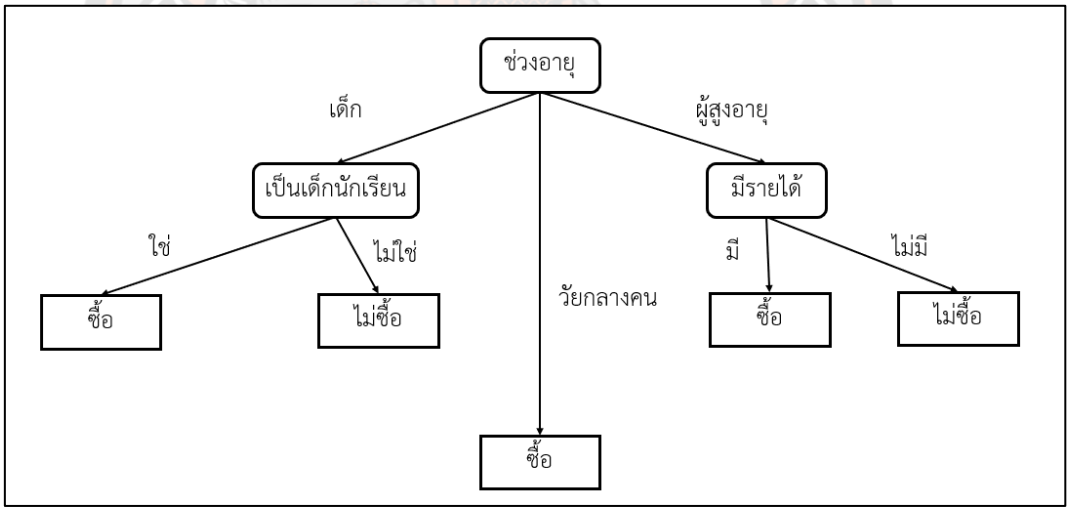
$$= \left(\frac{1}{3}\right)\left(\frac{3}{3}\right)(0.000914)\left(\frac{3}{8}\right)$$

$$= 0.000038$$

เนื่องจากความน่าจะเป็นของกลุ่ม Y_1 มีค่ามากกว่ากลุ่มที่ Y_2 ดังนั้นจะจำแนกให้ชุดข้อมูลใหม่ดังกล่าวอยู่กลุ่ม Y_1

2.2.3 ต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 (Decision Tree C4.5)

ต้นไม้ตัดสินใจเป็นหนึ่งในเทคนิคการจำแนกทางด้านการเรียนรู้ของเครื่องโดยใช้การคำนวณค่าความสัมพันธ์ระหว่างตัวแปรอิสระกับกลุ่มคำตอบ (Class Label) การสร้างต้นไม้จะสร้างจากบนลงล่าง (Top - Down) โดยตัวแบบการจำแนกที่ได้จะมีลักษณะคล้ายต้นไม้ซึ่งภายในต้นไม้จะมีส่วนประกอบดังนี้ โหนดราก (Root Node) โหนดภายใน (Internal Node) กิ่ง (Branch) และใบ (Leaf) (สายชล สินสมบูรณ์ทอง, 2560)



ภาพที่ 3 ตัวแบบต้นไม้ตัดสินใจ

จากภาพที่ 3 คือตัวแบบการจำแนกต้นไม้ตัดสินใจโดยมีคำตอบหรือตัวแปรตาม 2 ค่าประกอบด้วยซื้อและไม่ซื้อ มีตัวแปรอิสระประกอบด้วยเป็นนักเรียนและมีรายได้เป็นโหนดเพื่อใช้ในการตัดสินใจ การพยากรณ์การจำแนกกลุ่มจะพิจารณาจากบนลงล่างยกตัวอย่างเช่น ชุดข้อมูลใหม่มีตัวแปรอิสระคือ x=(ช่วงอายุ = ผู้สูงอายุ, มีรายได้ = มี, เป็นนักเรียน = ไม่ใช่) ดังนั้นจะพยากรณ์การจำแนกให้ชุดข้อมูลดังกล่าวเป็นลูกค้าอยู่กลุ่มที่จะซื้อผลิตภัณฑ์

อัลกอริทึม ID3 (Iterative Dichotomiser 3)

ID3 เป็นอัลกอริทึมแรกที่ใช้ในการสร้างต้นไม้ตัดสินใจถูกคิดค้นโดย John Ross Quinlan ซึ่งจะใช้แนวคิดที่ว่าตัวแปรอิสระตัวใดสามารถจำแนกชุดข้อมูลได้ดีที่สุด (Han et al., 2012) โดยในการเลือกตัวแปรอิสระมาเป็นโหนดของต้นไม้ตัดสินใจวัดจากค่าเกน ซึ่งคำนวณจากเกนสารสนเทศ (Information gain) โดยสมการที่ 17

$$Info(D) = -\sum_{i=1}^{m} p_i log_2(p_i) \quad (17)$$

โดยที่ D คือ ชุดข้อมูลที่สนใจ

p_i คือ ความน่าจะเป็นของจำนวนของตัวแปรตาม i ต่อจำนวนตัวแปรตามทั้งหมด

i คือ กลุ่มของตัวแปรตาม ซึ่งมีทั้งหมด m กลุ่ม

m คือ จำนวนกลุ่มทั้งหมดของตัวแปรตาม

จากนั้นก็จะหาเกนสารสนเทศของตัวแปรอิสระหรือตัวแปรอิสระแต่ละตัวโดยสมการที่ 18

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \quad (18)$$

โดยที่ D คือ ชุดข้อมูลที่สนใจ

D_j คือ ตัวแปรอิสระตัวที่ j

j คือ กลุ่มของตัวแปรอิสระ ซึ่งมีทั้งหมด v กลุ่ม

v คือ จำนวนกลุ่มทั้งหมดของตัวแปรอิสระ

เมื่อได้เกนสารสนเทศของข้อมูลทั้งหมดและเกนสารสนเทศของตัวแปรอิสระแล้ว ขั้นตอนต่อไปหาค่าเกนของตัวแปรอิสระแต่ละตัว โดยสมการที่ 19 จากนั้นจึงเลือกตัวแปรอิสระที่มีค่าเกนสูงที่สุดเป็นตัวจำแนกชุดข้อมูล

$$Gain(A) = Info(D) - Info_A(D) \quad (19)$$

โดยที่ D คือ ชุดข้อมูลที่สนใจ

A คือ ตัวแปรอิสระที่สนใจ

อัลกอริทึม C4.5

C4.5 เป็นอัลกอริทึมที่ใช้ในการสอนต้นไม้ตัดสินใจ พัฒนาต่อมาจากอัลกอริทึม ID3 หลักการทำงานของอัลกอริทึม C4.5 ใช้การคำนวณหาเกนสารสนเทศจากชุดข้อมูลทั้งหมดเช่นเดียวกับอัลกอริทึม ID3 เพื่อมาหาค่าเกนของตัวแปรอิสระแต่ละตัว เลือกตัวแปรที่มีค่าเกนสูงที่สุดเป็นโหนดราก แล้วจึงแตกกิ่งไปจนถึงใบ แต่จะมีส่วนที่แตกต่างจาก ID3 ตรงที่มีการแก้ไขความเอนเอียงของค่า

เกน โดยการปรับค่าเกนให้ถูกต้องจากการใช้ค่าสารสนเทศของการจำแนก (Split information) ของตัวแปรอิสระแต่ละตัว ดังสมการที่ 20 และเมื่อได้ค่าสารสนเทศของการจำแนกแล้ว สามารถคำนวณอัตราส่วนเกน (Gain ratio) เพื่อลดความเอนเอียงโดยสมการที่ 21

$$Split\ Info(A) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times log_2\left(\frac{|D_j|}{|D|}\right) \tag{20}$$

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)} \tag{21}$$

โดยที่ A คือ ตัวแปรอิสระที่สนใจ

D_j คือ ตัวแปรอิสระตัวที่ j

j คือ กลุ่มของตัวแปรอิสระ ซึ่งมีทั้งหมด v กลุ่ม

v คือ จำนวนกลุ่มทั้งหมดของตัวแปรอิสระ

หัวข้อถัดไปผู้วิจัยจะแสดงตัวอย่างการคำนวณด้วยเทคนิคต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 ด้วยชุดข้อมูลที่เป็นข้อมูลความเสี่ยงในการให้เครดิตดังตารางที่ 3

จากชุดข้อมูลในตารางที่ 3 เมื่อพิจารณาตัวแปรอิสระ Income พบว่าตัวแปรเป็นเชิงปริมาณ ซึ่งการทำงานภายใต้ตัวแปรอิสระเชิงปริมาณ ต้นไม้ตัดสินใจจะทำการแบ่งข้อมูลด้วยค่าคงที่ค่าหนึ่ง จากนั้นนำตัวแปรอิสระที่ทำการแปลงแล้วมาคำนวณหาค่าเกน ในที่นี้ขอยกตัวอย่างการแบ่งข้อมูลในตัวแปร Income ด้วยค่าคงที่ 50 และ 75 ดังนั้นจะได้ตัวแปรอิสระ Income ≤ 50 และ Income ≤ 75 ดังตารางต่อไปนี้

Customer	Saving	Assets	Income ≤ 50	Income ≤ 75	Credit Risk
1	Medium	High	No	Yes	Good
2	Low	Low	Yes	Yes	Bad
3	High	Low	Yes	Yes	Bad
4	Medium	High	Yes	Yes	Good
5	Low	High	No	No	Good
6	High	Low	Yes	Yes	Good
7	Low	Low	Yes	Yes	Bad
8	Medium	Low	No	Yes	Good

ขั้นตอนที่ 1 คำนวณค่าเกนสารสนเทศ โดยพิจารณาจากกลุ่มคำตอบของตัวแปรตาม

$$Info(D) = -\frac{3}{8}\log_2\left(\frac{3}{8}\right) - \frac{5}{8}\log_2\left(\frac{5}{8}\right) = 0.954$$

ขั้นตอนที่ 2 คำนวณเกนสารสนเทศของข้อมูลของทุกตัวแปรอิสระ

$$Info_{saving}(D) = \frac{3}{8}\left(-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right) + \frac{3}{8}\left(-\frac{3}{3}\log_2\left(\frac{3}{3}\right)\right) + \frac{2}{8}\left(-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)$$

$$= 0.594$$

$$Info_{Assets}(D) = \frac{5}{8}\left(-\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right) + \frac{3}{8}\left(-\frac{3}{3}\log_2\left(\frac{3}{3}\right)\right)$$

$$= 0.805$$

$$Info_{Income \le 50}(D) = \frac{3}{8}\left(-\frac{3}{3}\log_2\left(\frac{3}{3}\right)\right) + \frac{5}{8}\left(-\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right)$$

$$= 0.607$$

$$Info_{Income \le 75}(D) = \frac{1}{8}\left(-\frac{1}{1}\log_2\left(\frac{1}{1}\right)\right) + \frac{7}{8}\left(-\frac{4}{7}\log_2\left(\frac{3}{7}\right)\right)$$

$$= 0.611$$

ขั้นตอนที่ 3 คำนวณค่าเกนแปรอิสระ

$$Gain\ (Saving) = 0.954 - 0.594 = 0.36$$

$$Gain\ (Assets) = 0.954 - 0.805 = 0.149$$

$$Gain\ (Income \le 50) = 0.954 - 0.607 = 0.347$$

$$Gain\ (Income \le 75) = 0.954 - 0.611 = 0.343$$

ขั้นตอนที่ 4 คำนวณหาค่าสารสนเทศของการจำแนกของทุกตัวแปรอิสระ

$$Split\ Info\ (Saving) = -\left(\frac{3}{8}\right)\times\log_2\left(\frac{3}{8}\right) - \left(\frac{3}{8}\right)\times\log_2\left(\frac{3}{8}\right) - \left(\frac{2}{8}\right)\times\log_2\left(\frac{2}{8}\right)$$

$$= 1.781$$

$$Split\ Info\ (Assets) = -\left(\frac{5}{8}\right)\times\log_2\left(\frac{5}{8}\right)-\left(\frac{3}{8}\right)\times\log_2\left(\frac{3}{8}\right)$$

$$= 0.955$$

$$Split\ Info\ (Income \le 50) = -\left(\frac{5}{8}\right)\times\log_2\left(\frac{5}{8}\right)-\left(\frac{3}{8}\right)\times\log_2\left(\frac{3}{8}\right)$$

$$= 0.955$$

$$Split\ Info\ (Income \le 75) = -\left(\frac{1}{8}\right)\times\log_2\left(\frac{1}{8}\right)-\left(\frac{7}{8}\right)\times\log_2\left(\frac{7}{8}\right)$$

$$= 0.544$$

ขั้นตอนที่ 5 คำนวณหาค่าอัตราส่วนเกนของทุกตัวแปรอิสระ

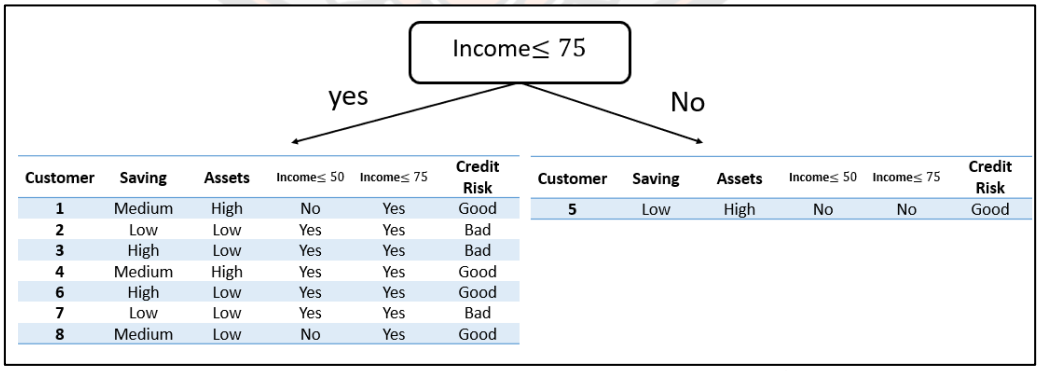
$$Gain\ Ratio\ (Saving) = \frac{0.36}{1.781} = 0.202$$

$$Gain\ Ratio\ (Assets) = \frac{0.149}{0.955} = 0.156$$

$$Gain\ Ratio\ (Income \le 50) = \frac{0.347}{0.955} = 0.363$$

$$Gain\ Ratio\ (Income \le 75) = \frac{0.343}{0.544} = 0.631$$

จากขั้นตอนที่ 3 ตัวแปรอิสระที่มีค่าอัตราส่วนเกนมากที่สุดคือ Income ≤ 75 ดังนั้นเราจะเลือก Income ≤ 75 เป็นโหนดเริ่มต้นหรือโหนดราก จะได้ตัวแบบการจำแนกเบื้องต้นดังภาพที่ 4

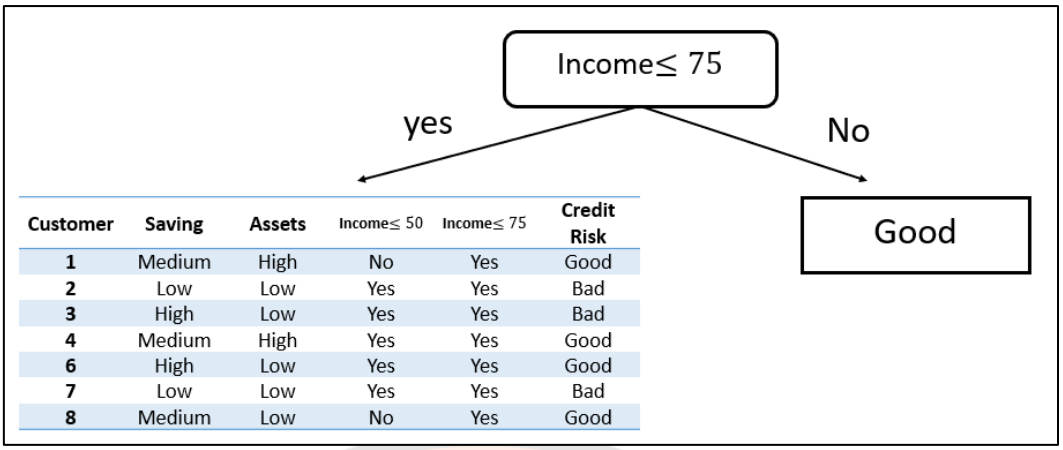


Customer	Saving	Assets	Income≤ 50	Income≤ 75	Credit Risk
1	Medium	High	No	Yes	Good
2	Low	Low	Yes	Yes	Bad
3	High	Low	Yes	Yes	Bad
4	Medium	High	Yes	Yes	Good
6	High	Low	Yes	Yes	Good
7	Low	Low	Yes	Yes	Bad
8	Medium	Low	No	Yes	Good

Customer	Saving	Assets	Income≤ 50	Income≤ 75	Credit Risk
5	Low	High	No	No	Good

ภาพที่ 4 เหตุการณ์การเลือกตัวแปรอิสระเป็นโหนดเริ่มต้น

จากภาพที่ 4 จะสังเกตได้ว่าชุดข้อมูลโหนดภายในจากกิ่ง No ข้อมูลจะเป็นคลาส Good ดังนั้นจะกำหนดให้เป็นใบ Good ดังภาพที่ 5

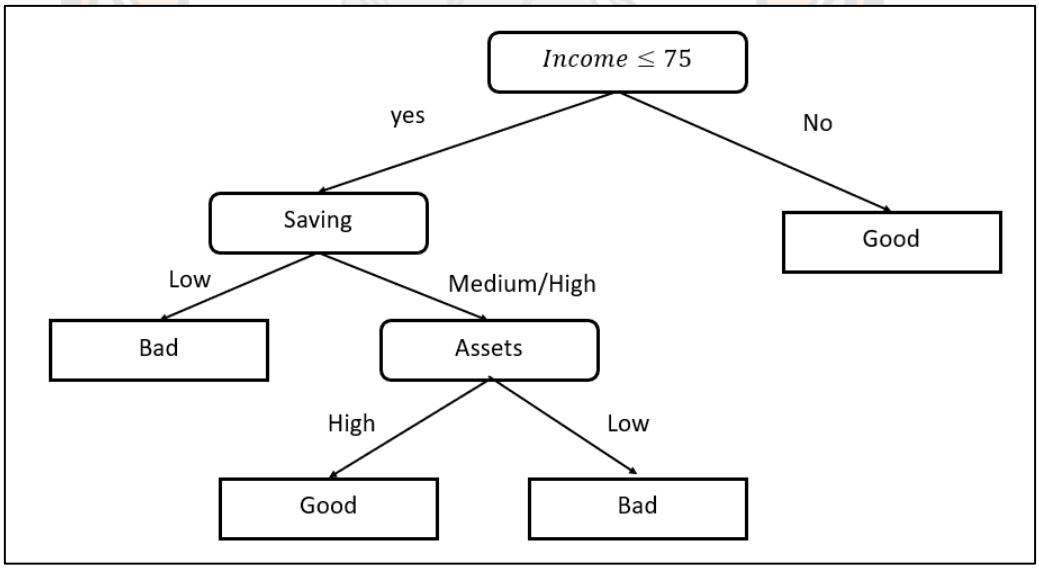


Customer	Saving	Assets	Income≤ 50	Income≤ 75	Credit Risk
1	Medium	High	No	Yes	Good
2	Low	Low	Yes	Yes	Bad
3	High	Low	Yes	Yes	Bad
4	Medium	High	Yes	Yes	Good
6	High	Low	Yes	Yes	Good
7	Low	Low	Yes	Yes	Bad
8	Medium	Low	No	Yes	Good

ภาพที่ 5 การพิจารณาโหนดภายใน yes

จากภาพที่ 5 โหนดรากคือ Income ≤75 ประกอบไปด้วยกิ่ง Yes และ No โดยกิ่ง No มีใบคือ Good ในขณะที่กิ่ง yes ต้องมีการพิจารณาโหนดภายในต่อไป

ขั้นตอนที่ 6 ทำซ้ำในขั้นตอนที่ 1-5 จนไม่สามารถแตกกิ่งได้อีกหรือถึงความลึกของต้นไม้ที่ผู้ศึกษากำหนด จากนั้นหยุดการคำนวณ

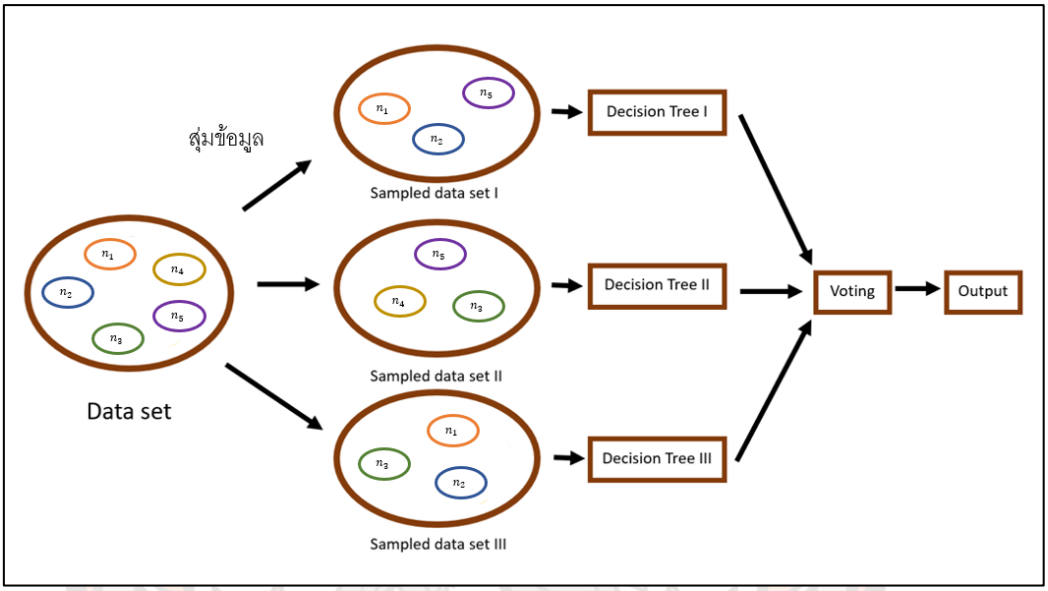


ภาพที่ 6 แสดงต้นไม้ตัดสินใจ ที่แตกกิ่งเสร็จสมบูรณ์

2.2.4 เทคนิคป่าสุ่ม (Random Forest)

เทคนิคป่าสุ่มเป็นหนึ่งในเทคนิคการจำแนกทางด้านการเรียนรู้ของเครื่องโดยมีหลักการทำงานคือการปลูกต้นไม้ตัดสินใจหลาย ๆ ต้นเพื่อเพิ่มประสิทธิภาพการจำแนกจากการปลูกต้นไม้เพียงต้นเดียว ในขั้นตอนการทำงานของเทคนิคป่าสุ่มจะเริ่มต้นจากการสุ่มชุดข้อมูลให้ต้นไม้แต่ละต้นและ

ทำการปลูกต้นไม้ตัดสินใจจากข้อมูลที่ได้รับ โดยการสร้างต้นไม้แต่ละต้นจะมีลักษณะแบบไม่ตัดแต่งกิ่ง (Unpruned) (วัชรีวรรณ จิตต์สกุล, 2560) จากนั้นให้ต้นไม้ตัดสินใจแต่ละต้นทำการพยากรณ์และทำการโหวตจากคำตอบที่ได้จากต้นไม้แต่ละต้น โดยจะเลือกคำตอบที่ได้รับการโหวตมากที่สุด (Robin & Jean-Michel, 2020) (พัฒนพงษ์ ดลรัตน์, 2560) (สห ธิติถามวัต, 2562) (Thanh Noi & Kappas, 2017)

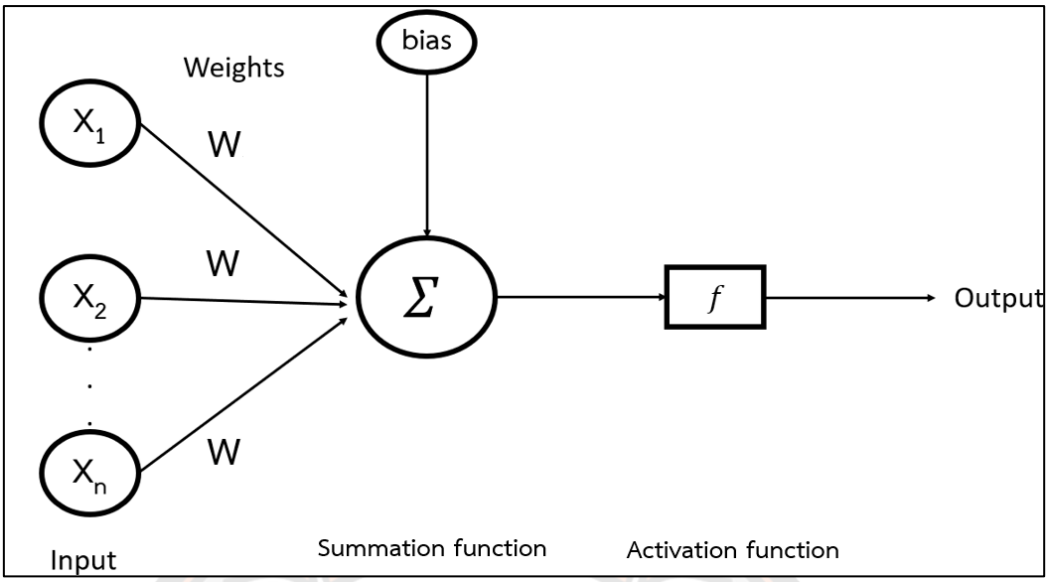


ภาพที่ 7 กระบวนการทำงานของเทคนิคป่าสุ่ม

ที่มา : https://medium.com/@witchapongdaroontham

2.2.5 โครงข่ายประสาทเทียม (Artificial Neural Network)

โครงข่ายประสาทเทียมเป็นหนึ่งในเทคนิคการจำแนกทางด้านการเรียนรู้ของเครื่องซึ่งมีกลไกการเรียนรู้ที่เลียนแบบการทำงานของระบบประสาทในสิ่งมีชีวิต โดยภายในจะประกอบด้วยโหนดจำลองมาจากรอยประสานประสาท (Synapse) ระหว่างใยประสาทนำเข้า (Dendrite) ของเซลล์ประสาทตัวหนึ่งและแกนประสาทนำออก (Axon) ของเซลล์ประสาทอีกตัวหนึ่ง โดยมีฟังก์ชันการแปลงเป็นตัวกำหนดสัญญาณส่งออก ในที่นี้เดนไดรต์ (Dendrites) เป็นตัวนำข้อมูลเข้าสู่เซลล์ประสาท ประสาทซึ่งเปรียบเทียบได้กับหน่วยข้อมูลเข้า การส่งข้อมูลออกจากโหนดผ่านทางเอกซอน (Axon) เปรียบเทียบได้กับหน่วยข้อมูลออกในโครงข่ายประสาทเทียม โดยโหนดในโครงข่ายประสาทเทียมจะมีการประมวลผล 2 ขั้นตอนคือ ขั้นตอนการหาผลรวมและขั้นตอนการแปลง โดยโครงข่ายประสาทเทียมจะมีการป้อนข้อมูลเข้าและการกำหนดค่าน้ำหนักต่าง ๆ (อกนิษฐ์ ทองจิตร, 2562)



ภาพที่ 8 แบบจำลองการทำงานของโครงข่ายประสาทเทียมอย่างง่าย

จากภาพที่ 8 คือแบบจำลองการทำงานของโครงข่ายประสาทเทียมอย่างง่าย ประกอบด้วยโหนดเชื่อมโยงกันหลาย ๆ ตัว คล้ายคลึงกับการเชื่อมต่อกันของเซลล์ประสาทแต่ละโหนดจะรับข้อมูลนำเข้า (Input) เข้ามาโดยจะรวบรวมโดยใช้ฟังก์ชันผลรวม จากนั้นนำผลลัพธ์ที่จากฟังก์ชันผลรวมนำมาคำนวณโดยใช้ฟังก์ชันกระตุ้น โดยผลลัพธ์หลักจากนำเข้าฟังก์ชันกระตุ้นจะเป็นผลลัพธ์ (Output) เพื่อส่งไปยังโหนดถัดไป จากภาพที่ 8 มีรายละเอียดดังต่อไปนี้

1. ข้อมูลนำเข้า (Input) ในที่นี้ได้แก่ x_1 x_2 และ x_n

2. โหนด (Node) ทำหน้าที่รับข้อมูลเข้ามาในโหนด โดยมีการกำหนดค่าน้ำหนักของข้อมูลนำเข้า โหนดจะประมวลผลรวมและทำการแปลงค่าผ่านฟังก์ชันกระตุ้น

3. ผลลัพธ์ (Output) ทำหน้าที่ส่งออกข้อมูลที่ผ่านการประมวลผลจากโหนดดังสมการ 22 และ 23

$$net_j = \sum_i W_{ij} x_{ij} \tag{22}$$

$$= W_{0j} x_{0j} + W_{1j} x_{1j} + \ldots + W_{nj} x_{nj}$$

$$f(net_j) = \frac{1}{1+e^{-net_j}} \tag{23}$$

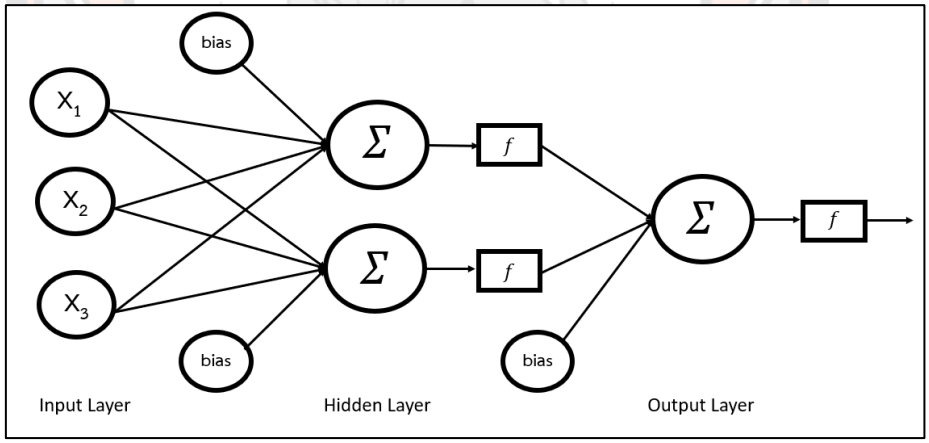
เมื่อ x_{ij} คือ ข้อมูลนำเข้าที่ i ไปยังโหนดที่ j

W_{ij} คือ น้ำหนักคูณกับข้อมูลนำเข้าที่ i ไปยังโหนดที่ j

จากสมการที่ 22 คือฟังก์ชันผลรวม (Summation function) เป็นผลรวมของข้อมูลนำเข้าที่ถ่วงด้วยค่าน้ำหนัก (W)

จากสมการที่ 23 คือฟังก์ชันกระตุ้น (Activation function) เป็นส่วนที่ทำหน้าที่แปลงค่าที่ได้จากฟังก์ชันผลรวมก่อนส่งผลลัพธ์ไปยังโหนดถัดไป ในงานวิจัยนี้จะใช้ฟังก์ชันซิกมอยด์ (Sigmoid function) ซึ่งฟังก์ชันการแปลงซิกมอยด์จะทำให้ช่วงข้อมูลให้อยู่ในช่วง 0 ถึง 1

เมื่อนำโครงข่ายประสาทเทียมอย่างง่ายจากภาพที่ 8 มาเชื่อมต่อกันหลาย ๆ โหนดเข้าด้วยกันทำให้เกิดเป็นลักษณะโครงข่ายเป็นชั้น ๆ (Layer) โดยโครงข่ายประสาทเทียมประกอบไปด้วยชั้นนำเข้า (Input Layer) ชั้นซ่อน (Hidden Layer) และชั้นผลลัพธ์ (Output Layer) ซึ่งหน้าที่ของชั้นนำเข้าคือรับข้อมูลเข้ามาโดยจะไม่มีการคำนวณใด ๆ และส่งค่าผลลัพธ์ไปให้โหนดในชั้นซ่อน หน้าที่ของชั้นซ่อนคือรับข้อมูลนำเข้าจากชั้นนำเข้า (โดยผลลัพธ์ที่ส่งมาจากชั้นนำเข้าจะกลายเป็นข้อมูลนำเข้าในชั้นซ่อน) โดยจะรวบรวมข้อมูลนำเข้าด้วยฟังก์ชันผลรวมและส่งผลลัพธ์ผ่านฟังก์ชันกระตุ้นไปยังโหนดในชั้นผลลัพธ์ หน้าที่ของชั้นผลลัพธ์คือรับข้อมูลนำเข้าจากชั้นซ่อน (โดยผลลัพธ์ที่ส่งมาจากชั้นซ่อนจะกลายเป็นข้อมูลนำเข้าในชั้นผลลัพธ์) รวบรวมข้อมูลนำเข้าด้วยฟังก์ชันผลรวมและส่งผลลัพธ์ด้วยฟังก์กระตุ้น นำค่าผลลัพธ์ที่ได้จากชั้นผลลัพธ์มาพิจารณาเพื่อใช้ในการจำแนกกลุ่ม ซึ่งโหนดในโครงข่ายประสาทเทียมแต่ละตัวที่อยู่ในชั้นเดียวกันจะไม่มีการเชื่อมต่อกัน จะได้โครงข่ายประสาทเทียมดังภาพที่ 9



ภาพที่ 9 แบบจำลองการทำงานของโครงข่ายประสาทเทียมอย่างง่าย เชื่อมกันหลาย ๆ โหนดเข้าด้วยกัน

ในหัวข้อถัดไปผู้วิจัยจะแสดงตัวอย่างการคำนวณการจำแนกด้วยเทคนิคโครงข่ายประสาทเทียม ด้วยชุดข้อมูลความเสี่ยงในการให้เครดิตจากตารางที่ 3

ขั้นตอนที่ 1 แปลงชุดข้อมูลให้อยู่ในรูปการทำให้เป็นปรกติ โดยมีจะค่าอยู่ระหว่าง 0 ถึง 1 ดังสมการที่ 23

$$x^{*} = \frac{x - min(x)}{max(x) - min(x)} \quad (23)$$

โดยที่ x^{*} คือ ค่าที่ได้จากการแปลงค่าอยู่ในรูปการทำให้เป็นปรกติ

x คือ ค่าของตัวแปรอิสระ

$min(x)$ คือ ค่าของข้อมูลที่มีค่าน้อยที่สุดในชุดข้อมูล

$max(x)$ คือ ค่าของข้อมูลที่มีค่ามากที่สุดในชุดข้อมูล

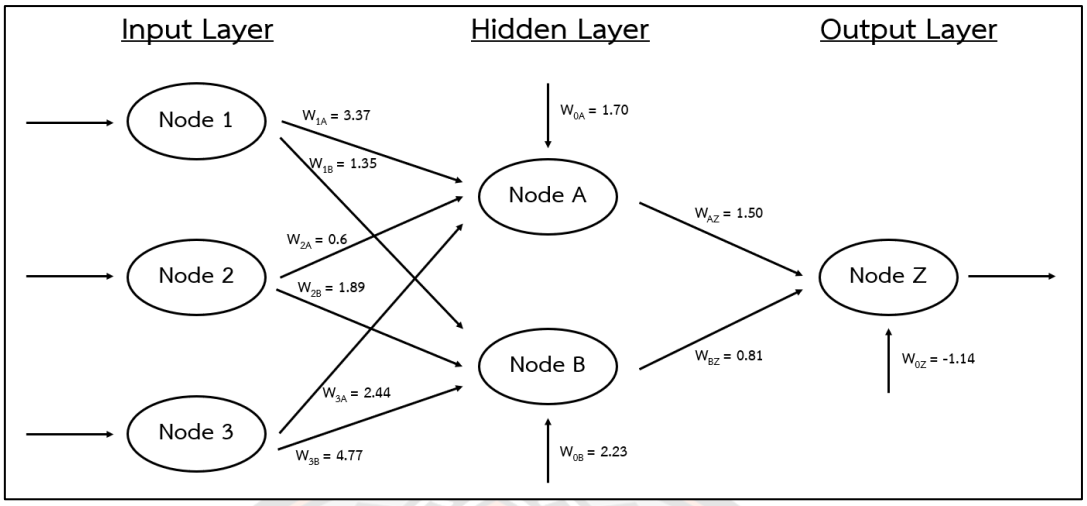
หลังจากทำการแปลงข้อมูลให้อยู่ในการทำให้เป็นปรกติจะได้ชุดข้อมูลดังตารางต่อไปนี้

Customer	Saving	Assets	Income	Credit Risk
1	0.5	1	0.667	1
2	0	0	0.333	0
3	1	0	0.000	0
4	0.5	1	0.333	1
5	0	1	1.000	1
6	1	0	0.000	1
7	0	0	0.000	0
8	0.5	0	0.667	1

ขั้นตอนที่ 2 นำชุดข้อมูลเข้าโปรแกรม โดยใช้แพ็คเกจ neuralnet จากโปรแกรม RStudio Version 1.4.1717 จะได้ค่าน้ำหนักถ่วงเริ่มต้น ได้ดังตารางต่อไปนี้

$x_0 = 1$	$W_{0A} = 1.70$	$W_{0B} = 2.23$	$W_{0Z} = -1.14$
$x_1 = 0.5$	$W_{1A} = 3.37$	$W_{1B} = 1.35$	$W_{AZ} = 1.50$
$x_2 = 1$	$W_{2A} = 0.6$	$W_{2B} = 1.89$	$W_{BZ} = 0.81$
$x_3 = 0.667$	$W_{3A} = 2.44$	$W_{3B} = 4.77$	

จากตารางน้ำหนักถ่วงเริ่มต้นข้างต้น นำมาสร้างตัวแบบการจำแนกโครงข่ายประสาทเทียมได้ดังภาพที่ 10



ภาพที่ 10 แบบจำลองการทำงานของโครงข่ายประสาท
ดัดแปลงมาจากโปรแกรม Weka Version 3.8.5

ขั้นตอนที่ 3 คำนวณโหนด A ในชั้นซ่อน (Hidden Layer) ในที่นี้จะขอยกตัวอย่างชุดข้อมูลลูกค้าคนที่ 1 (Customer = 1) ซึ่งมีค่าของตัวแปรอิสระดังนี้ Saving = 0.5, Assets = 1, Income = 0.667 สามารถคำนวณได้ดังนี้

$$net_A = \sum_i W_{iA} x_{iA}$$

$$= W_{0A}(1) + W_{1A}x_{1A} + W_{2A}x_{2A} + W_{3A}x_{3A}$$

$$= 1.70 + (3.37)(0.5) + (0.6)(1) + (2.44)(0.667)$$

$$= 5.613$$

$$f(net_A) = \frac{1}{1+e^{-net_A}}$$

$$= \frac{1}{1+e^{-5.613}}$$

$$= 0.9964$$

คำนวณโหนด B ในชั้นซ่อน

$$net_B = \sum_i W_{iB} x_{iB}$$

$$= W_{0B}(1) + W_{1B}x_{1B} + W_{2B}x_{2B} + W_{3B}x_{3B}$$

$$= 2.23 + (1.35)(0.5) + (1)(1.89) + (0.667)(4.77)$$

$$= 7.977$$

$$f(net_B) = \frac{1}{1+e^{-net_B}}$$
$$= \frac{1}{1+e^{-7.977}}$$
$$= 0.9992$$

ขั้นตอนที่ 4 คำนวณโหนด Z ในชั้นผลลัพธ์

$$net_Z = \sum_i W_{iZ} x_{iZ}$$
$$= W_{0Z}(1) + W_{AZ} x_{AZ} + W_{BZ} x_{BZ}$$
$$= -1.14 + (1.50)(0.9964) + (0.81)(0.9992)$$
$$= 1.164$$

$$f(net_Z) = \frac{1}{1+e^{-net_Z}}$$
$$= \frac{1}{1+e^{-1.164}}$$
$$= 0.762$$

เนื่องจากชุดข้อมูลที่ 1 ให้ค่าโหนด Z ในชั้นผลลัพธ์ค่าเท่ากับ 0.762 ซึ่งมากกว่า 0.5 ดังนั้นพยากรณ์ว่าข้อมูลอยู่กลุ่มที่ 1 นั่นคือพยากรณ์ว่า Credit Risk อยู่กลุ่ม Good จากนั้นนำข้อมูลชุดที่เหลือมาคำนวณเพื่อพยากรณ์การจำแนก

การปรับปรุงน้ำหนักถ่วงด้วยการทำงานแบบแพร่ย้อนกลับ

การปรับปรุงน้ำหนักถ่วงด้วยการทำงานแบบแพร่ย้อนกลับ เป็นหนึ่งในกระบวนการของโครงข่ายประสาทเทียมโดยมีจุดประสงค์เพื่อพัฒนาประสิทธิภาพการจำแนกของน้ำหนักถ่วงชุดเดิม โดยจะมีการคำนวณเพื่อหาน้ำหนักถ่วงที่ปรับใหม่ทุก ๆ ชุดข้อมูล ทำเช่นนี้จนครบทุกชุดข้อมูลนับเป็น 1 รอบ (Epoch) ซึ่งการทำซ้ำหลาย ๆ รอบ จะทำให้ความคลาดเคลื่อนระหว่างค่าจริงกับค่าพยากรณ์ลดลงด้วยการเคลื่อนลงตามความชัน (Gradient Descent Method)

$$\nabla SSE(w) = \left[\frac{\partial SSE}{\partial w_1}, \frac{\partial SSE}{\partial w_2}, \ldots, \frac{\partial SSE}{\partial w_n} \right] \quad (24)$$

การหาน้ำหนักถ่วงด้วยการทำงานแบบการส่งค่าย้อนกลับ มีพารามิเตอร์ที่เกี่ยวข้อง ได้แก่ อัตราการเรียนรู้ (Learning rate : η) และค่าโมเมนตัม (Momentum : α) ที่มีค่าอยู่ในช่วง 0 ถึง 1 โดยการปรับน้ำหนักถ่วงสามารถคำนวณดังสมการที่ 25

$$w_{ij,new} = w_{ij,\text{current}} + \Delta w_{ij,\text{current}} + \alpha \Delta w_{ij,\text{previous}} \quad (25)$$

เมื่อ $w_{ij,\text{current}} = \eta\delta_j x_{ij}$

$w_{ij,new}$ คือ น้ำหนักถ่วงใหม่ของตัวแปรอิสระที่ i ไปยังโหนด j

$w_{ij,\text{current}}$ คือ น้ำหนักถ่วงเดิมของตัวแปรอิสระที่ I ไปยังโหนด j

η คือ อัตราการเรียนรู้

x_{ij} คือ ค่าตัวแปรอิสระที่ i ไปยังโหนด j

α คือ ค่าโมเมนตัม

$\Delta w_{ij,\text{previous}}$ คือ $\Delta w_{ij,\text{current}}$ ของจุดทดลองก่อนหน้า

δ_j คือ ค่าความคลาดเคลื่อนของโหนด

$\delta_j = output_j(1- output_j)(actual_j - output_j)$ เมื่อโหนด j อยู่ในชั้นผลลัพธ์ และ

$\delta_j = output_j(1 - output_j)\sum_{downstream} w_{jk}\delta_j$ เมื่อโหนด j อยู่ในชั้นซ่อน และโหนด k อยู่ในชั้นถัดไปทางขวา

ในหัวข้อถัดไปผู้วิจัยจะแสดงตัวอย่างการคำนวณการปรับปรุงน้ำหนักถ่วงด้วยการทำงานแบบแพร่ย้อนกลับด้วยชุดข้อมูลและน้ำหนักถ่วงจากตัวอย่างก่อนหน้า ดังนี้

ขั้นตอนที่ 1 เริ่มต้นจากการกำหนดพารามิเตอร์ที่เกี่ยวข้องประกอบด้วย อัตราการเรียนรู้เท่ากับ 0.2 และค่าโมเมนตัมเท่ากับ 0.5 ในที่นี้จะใช้ชุดข้อมูลลูกค้าคนที่ 1 (Customer = 1) ซึ่งมีค่าของตัวแปรอิสระดังนี้ Saving = 0.5, Assets = 1, Income = 0.50 สามารถคำนวณได้ดังนี้

$\delta_z = output_z(1 - output_z)(actual_z - output_z)$

$= (0.762)(1-0.762)(1-0.762)$

$= 0.0432$

$\delta_A = output_A(1 - output_A)\sum_{downstream} w_{jk}\delta_j$

$= output_A(1 - output_A)\, w_{AZ}\, \delta_Z$

$= (0.9964)(1-0.9964)(1.5)(0.0432)$

$= 0.0002324402$

$\delta_B = output_B(1 - output_B)\sum_{downstream} w_{jk}\delta_j$

$= output_B(1 - output_B)\, w_{BZ}\, \delta_Z$

$= (0.9992)(1-0.9992)(0.81)(0.0432)$

$= 0.00002797121$

เนื่องจากครั้งนี้เป็นการปรับปรุงน้ำหนักถ่วงครั้งแรก ดังนั้น

$\alpha\Delta w_{0Z,\text{previous}} = (0.5)(0) = 0$

พิจารณาปรับปรุงค่าน้ำหนักถ่วง W_{0Z} ได้เป็น

$$W_{0Z} = w_{0Z,current} + \Delta w_{0Z,current} + \alpha \Delta w_{0Z,previous}$$

$$= w_{0Z,current} + \eta \delta_Z (1) + \alpha \Delta w_{0Z,previous}$$

$$= -1.14 + (0.2)(0.0432) + 0 = -1.13$$

พิจารณาปรับปรุงค่าน้ำหนักถ่วงในชั้นซ่อน W_{AZ}

$$W_{AZ} = w_{AZ,current} + \Delta w_{AZ,current} + \alpha \Delta w_{AZ,previous}$$

$$= w_{AZ,current} + \eta \delta_Z output_A + \alpha \Delta w_{AZ,previous}$$

$$= 1.50 + (0.2)(0.0432)(0.9964)$$

$$= 1.508609$$

พิจารณาปรับปรุงค่าน้ำหนักถ่วง W_{BZ} ได้เป็น

$$W_{BZ} = w_{BZ,current} + \Delta w_{BZ,current} + \alpha \Delta w_{BZ,previous}$$

$$= w_{BZ,current} + \eta \delta_Z output_B + \alpha \Delta w_{BZ,previous}$$

$$= 0.81 + (0.2)(0.0432)(0.992) + 0$$

$$= 0.808633$$

พิจารณาปรับปรุงค่าน้ำหนักถ่วงในชั้นซ่อน W_{0A}

$$W_{0A} = w_{0A,current} + \Delta w_{0A,current} + \alpha \Delta w_{0A,previous}$$

$$= w_{0A,current} + \eta \delta_A (1) + \alpha \Delta w_{0A,previous}$$

$$= 1.70 + (0.2)(0.0002324402) + 0$$

$$= 1.700046$$

พิจารณาปรับปรุงค่าน้ำหนักถ่วงในชั้นซ่อน W_{1A}

$$W_{1A} = w_{1A,current} + \Delta w_{1A,current} + \alpha \Delta w_{1A,previous}$$

$$= w_{1A,current} + \eta \delta_A x_1 + \alpha \Delta w_{1A,previous}$$

$$= 3.37 + (0.2)(0.0002324402)(0.5) + 0$$

$$= 3.370023$$

พิจารณาปรับปรุงค่าน้ำหนักถ่วงในชั้นซ่อน W_{1B}

$$W_{1B} = w_{1B,current} + \Delta w_{1B,current} + \alpha \Delta w_{1B,previous}$$

$$= w_{1B,current} + \eta \delta_B x_1 + \alpha \Delta w_{1B,previous}$$

$$= 1.35 + (0.2)(0.00002797121)(0.5) + 0$$

$$= 1.350003$$

พิจารณาปรับปรุงค่าน้ำหนักถ่วงในชั้นซ่อน W_{2A}

$W_{2A} = w_{2A,current} + \Delta w_{2A,current} + \alpha \Delta w_{2A,previous}$

$= w_{2A,current} + \eta \delta_A x_2 + \alpha \Delta w_{2A,previous}$

= 0.6 + (0.2) (0.0002324402) (1) + 0

= 0.600232

พิจารณาปรับปรุงค่าน้ำหนักถ่วงในชั้นซ่อน W_{2B}

$W_{2B} = w_{2B,current} + \Delta w_{2B,current} + \alpha \Delta w_{2B,previous}$

$= w_{2B,current} + \eta \delta_B x_2 + \alpha \Delta w_{2B,previous}$

= 1.89 + (0.2) (0.00002797121) (1) + 0

= 1.890006

พิจารณาปรับปรุงค่าน้ำหนักถ่วงในชั้นซ่อน W_{3A}

$W_{3A} = w_{3A,current} + \Delta w_{3A,current} + \alpha \Delta w_{3A,previous}$

$= w_{3A,current} + \eta \delta_A x_3 + \alpha \Delta w_{3A,previous}$

= 2.44 + (0.2) (0.0002324402) (0.667) + 0

= 2.440031

พิจารณาปรับปรุงค่าน้ำหนักถ่วงในชั้นซ่อน W_{3B}

$W_{3B} = w_{3B,current} + \Delta w_{3B,current} + \alpha \Delta w_{3B,previous}$

$= w_{3B,current} + \eta \delta_B x_3 + \alpha \Delta w_{3B,previous}$

= 4.77+ (0.2) (0.00002797121) (0.667) + 0

= 4.770004

พิจารณาปรับปรุงค่าน้ำหนักถ่วงในชั้นซ่อน W_{0B}

$W_{0B} = w_{0B,current} + \Delta w_{0B,current} + \alpha \Delta w_{0B,previous}$

$= w_{0B,current} + \eta \delta_B (1) + \alpha \Delta w_{0B,previous}$

= 2.23 + (0.2) (0.00002797121) (1) + 0

= 2.230006

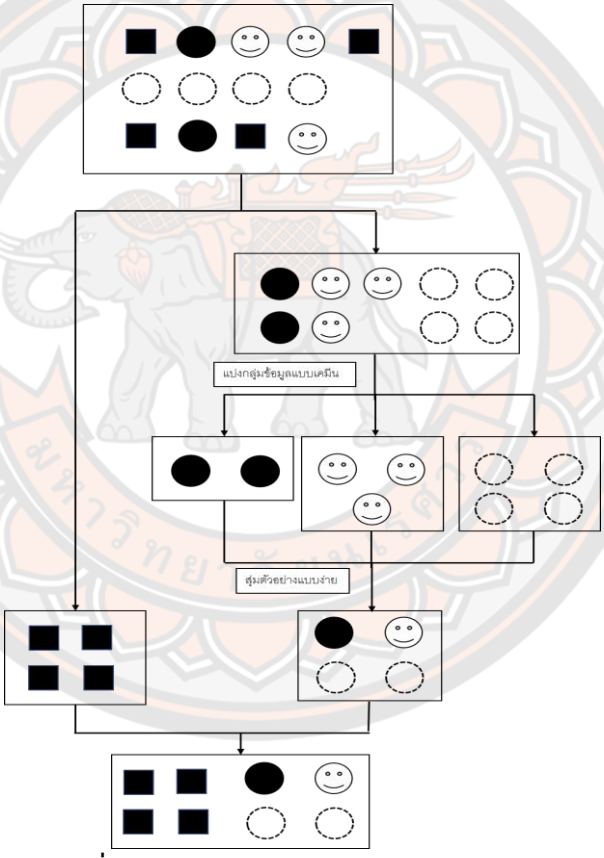
และทำเช่นนี้กับชุดข้อมูลที่เหลือ

2.3 การแบ่งกลุ่มข้อมูลแบบเคมีน (k-means)

ในการศึกษานี้ผู้วิจัยจะนำเสนอเทคนิคการวิเคราะห์แบ่งกลุ่ม (Cluster Analysis) มาประยุกต์ใช้เพื่อปรับปรุงข้อมูลให้มีความสมดุลซึ่งหลักการทำงานของการวิเคราะห์แบ่งกลุ่มคือการแบ่งกลุ่มข้อมูลออกเป็นกลุ่ม ๆ ตามคุณลักษณะของตัวแปรอิสระที่รวบรวมมา โดยมุ่งเน้นการจัดกลุ่ม

ของข้อมูลที่อยู่ในกลุ่มเดียวกันมีความคล้ายคลึงกันมากที่สุด และข้อมูลที่อยู่ต่างกลุ่มกันจะต้องมีความคล้ายคลึงกันน้อยที่สุด โดยเทคนิคที่ผู้วิจัยจะนำมาประยุกต์ใช้ได้แก่ การแบ่งกลุ่มข้อมูลแบบเคมีน

การแบ่งกลุ่มข้อมูลแบบเคมีน คือเทคนิคการแบ่งกลุ่มข้อมูลที่ได้รับความนิยมอย่างมากในการลดมิติของข้อมูล โดยมีหลักการทำงานคือใช้ระยะห่างเพื่อแบ่งกลุ่มข้อมูลจำนวน n สิ่งออกเป็น k กลุ่ม ซึ่งข้อมูลใดที่มีระยะห่างจากจุดศูนย์กลางของกลุ่มใดน้อยที่สุดจะจัดให้อยู่กลุ่มนั้น ในการศึกษานี้ผู้วิจัยจะนำการแบ่งกลุ่มข้อมูลแบบเคมีนเพื่อมาประยุกต์ใช้ในการปรังปรุงชุดข้อมูลอสมดุลให้สมดุล โดยมีแนวคิดคือนำกลุ่มที่มีสมาชิกมากมาวิเคราะห์ด้วยการแบ่งกลุ่มข้อมูลด้วยเทคนิคเคมีน จากนั้นทำการสุ่มข้อมูลเพื่อเป็นตัวแทนด้วยการสุ่มตัวอย่างแบบง่าย ดังภาพที่ 11



ภาพที่ 11 การปรับปรุงชุดข้อมูลอสมดุลให้สมดุล
ด้วยเทคนิคการแบ่งกลุ่มข้อมูลแบบเคมีน

งานวิจัยนี้จะใช้ระยะห่างยุคลิด (Euclidean distance: d) เพื่อใช้ในการแบ่งกลุ่มข้อมูล โดยมีสมการที่ 26 (สายชล สินสมบูรณ์ทอง, 2560)

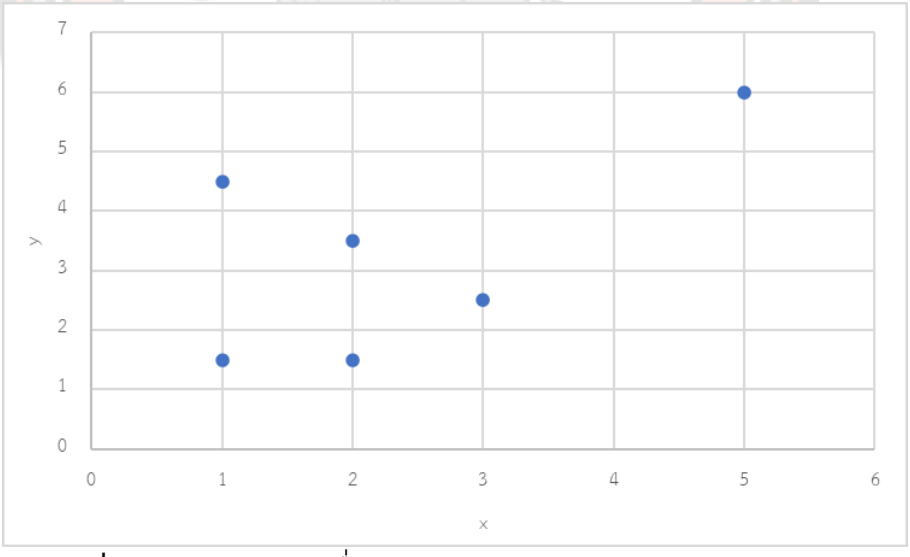
$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{26}$$

ในหัวข้อต่อไปผู้วิจัยจะแสดงตัวอย่างการคำนวณด้วยวิธีการแบ่งกลุ่มข้อมูลแบบเคมีน ด้วยชุดข้อมูลตัวอย่างดังตารางที่ 4

ตารางที่ 4 ข้อมูลตัวอย่างเพื่อแสดงการทำงานการแบ่งกลุ่มข้อมูลแบบเคมีน

ID	x	y
1	1	1.5
2	1	4.5
3	2	1.5
4	2	3.5
5	3	2.5
6	5	6

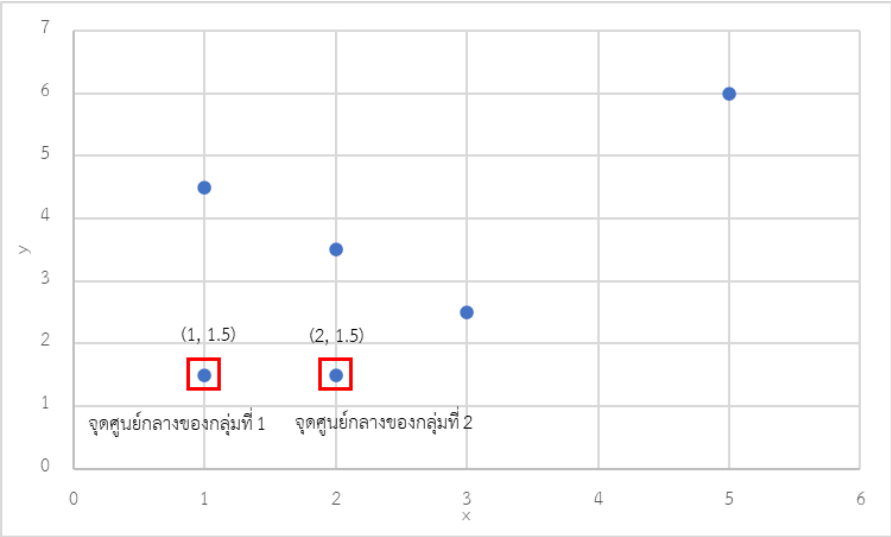
จากตารางที่ 4 นำมาสร้างกราฟได้ดังภาพที่ 12



ภาพที่ 12 ข้อมูลตัวอย่างเพื่อแสดงการทำงานการแบ่งกลุ่มข้อมูลแบบเคมีน

ขั้นตอนที่ 1 ทำการกำหนดจำนวนกลุ่ม ในที่นี้ผู้วิจัยกำหนดจำนวนกลุ่มเท่ากับ 2 (k=2)

ขั้นตอนที่ 2 ทำการสุ่มเลือกจุดศูนย์กลางของแต่ละกลุ่ม (Centroid) ในทีนี้จะสุ่มเลือกให้จุดศูนย์กลางของกลุ่มที่ 1 คือ {x=1, y=1.5} และจุดศูนย์กลางของกลุ่มที่ 2 คือ {x=2, y=1.5} สามารถแสดงดังภาพที่ 13



ภาพที่ 13 การสุ่มเลือกจุดศูนย์กลางของเทคนิคการแบ่งกลุ่มข้อมูลแบบเคมีน

ขั้นตอนที่ 3 คำนวณหาระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางของกลุ่มที่ 1 และจุดศูนย์กลางของกลุ่มที่ 2 ได้ดังต่อไปนี้

ID	จุดศูนย์กลางของกลุ่มที่ 1 {x=1, y=1.5}	จุดศูนย์กลางของกลุ่มที่ 2 {x=2, y=1.5}
1	$\sqrt{(1-1)^2+(1.5-1.5)^2}=0$	$\sqrt{(1-2)^2+(1.5-1.5)^2}=1$
2	$\sqrt{(1-1)^2+(4.5-1.5)^2}=3$	$\sqrt{(1-2)^2+(4.5-1.5)^2}=3.162$
3	$\sqrt{(2-1)^2+(1.5-1.5)^2}=1$	$\sqrt{(2-2)^2+(1.5-1.5)^2}=0$
4	$\sqrt{(2-1)^2+(3.5-1.5)^2}=2.2$	$\sqrt{(2-2)^2+(3.5-1.5)^2}=2$
5	$\sqrt{(3-1)^2+(2.5-1.5)^2}=2.236$	$\sqrt{(3-2)^2+(2.5-1.5)^2}=1.414$
6	$\sqrt{(5-1)^2+(6-1.5)^2}=6.021$	$\sqrt{(5-2)^2+(6-1.5)^2}=5.408$

ขั้นตอนที่ 4 พิจารณาการแบ่งกลุ่มของแต่ละชุดข้อมูลจากระยะห่างยุคลิดที่คำนวณได้ ยกตัวอย่างเช่น พิจารณาชุดข้อมูลที่ 2 (ID=2) พบว่าจะแบ่งกลุ่มให้ชุดข้อมูลดังกล่าวอยู่กลุ่มที่ 1 เนื่องจากระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางของกลุ่มที่ 1 (d=3) มีค่าน้อยกว่าระยะห่างไปยังจุดศูนย์กลางของกลุ่มที่ 2 (d=3.162) ในทำนองเดียวกันพิจารณาทุกชุดข้อมูลเพื่อแบ่งกลุ่ม

ในกรณีที่ระยะห่างของทั้ง 2 กลุ่มมีขนาดเท่ากันจะแบ่งให้ชุดข้อมูลอยู่กลุ่มใดกลุ่มใดกลุ่มหนึ่งก็ได้แต่ห้ามอยู่ทั้ง 2 กลุ่ม

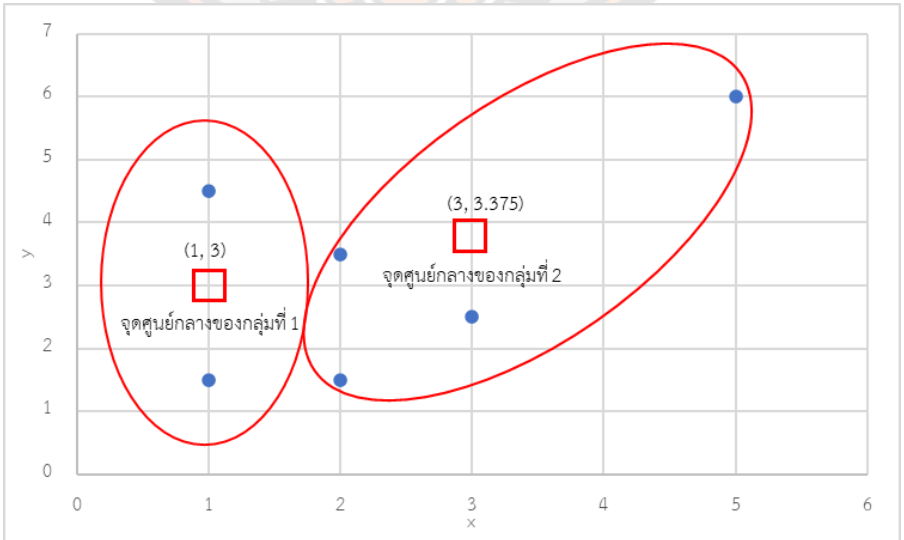
เมื่อพิจารณาระยะห่างยุคลิดเพื่อแบ่งกลุ่มครบทุกชุดข้อมูลแล้วจะได้ดังตารางต่อไปนี้

ID	x	y	จำแนกให้อยู่กลุ่มที่
1	1	1.5	1
2	1	4.5	1
3	2	1.5	2
4	2	3.5	2
5	3	2.5	2
6	5	6	2

ขั้นตอนที่ 5 หลังจากพิจารณาแบ่งกลุ่มครบทุกชุดข้อมูล จากนั้นคำนวณจุดศูนย์กลางของแต่ละกลุ่มใหม่ ดังนี้

	x	y
จุดศูนย์กลางของกลุ่มที่ 1	$\frac{(1+1)}{2}=1$	$\frac{(1.5+4.5)}{2}=3$
จุดศูนย์กลางของกลุ่มที่ 2	$\frac{(2+2+3+5)}{4}=3$	$\frac{(1.5+3.5+2.5+6)}{4}=3.375$

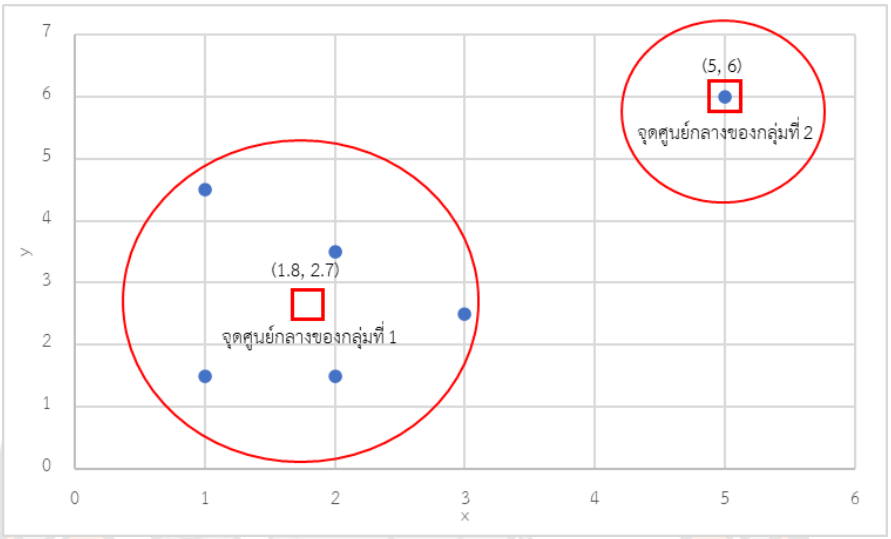
ดังนั้น จุดศูนย์กลางของกลุ่มที่ 1 คือ {x=1, y=3} และจุดศูนย์กลางของกลุ่มที่ 2 คือ {x=3, y=3.375} แสดงจุดศูนย์กลางของกลุ่มใหม่ดังภาพที่ 14



ภาพที่ 14 การพิจารณาเปลี่ยนจุดศูนย์กลางใหม่ของเทคนิคการแบ่งกลุ่มข้อมูลแบบเคมีน

ขั้นตอนที่ 6 ทำซ้ำในขั้นตอนที่ 1-5 จนกว่าจุดศูนย์กลางของแต่ละกลุ่มจะไม่เปลี่ยน

เมื่อคำนวณต่อไปเรื่อย ๆ จนจุดศูนย์กลางของแต่ละกลุ่มจะไม่เปลี่ยนจะได้จุดศูนย์กลางของกลุ่มที่ 1 คือ {x=1.8, y=2.7} และจุดศูนย์กลางของกลุ่มที่ 2 คือ {x=6, y=6} และจำแนกกลุ่มได้ดังภาพที่ 15



ภาพที่ 15 ตัวอย่างการแบ่งกลุ่มข้อมูล โดยใช้เทคนิคการแบ่งกลุ่มข้อมูลแบบเคมีน

ขั้นตอนที่ 7 คำนวณค่าผลรวมระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางของกลุ่ม (Within Groups Sum of Squares : WGSS) คำนวณดังสมการที่ 27

$$WGSS = \sum_{i=1}^{n_c} \sum_{x \in i} d\left(x, \overline{x}_i\right)^2 \qquad (27)$$

โดย x แทน ค่าของข้อมูล

$\overline{x}_i$ แทน จุดศูนย์กลางของกลุ่มที่ i เมื่อ i = 1, 2, ..., k

เนื่องจากตัวอย่างข้างต้นกำหนด k = 2 ดังนั้น

$$WGSS_1 = (1\text{-}1.8)^2 + (1.5\text{-}2.7)^2 + (1\text{-}1.8)^2 + (4.5\text{–}2.7)^2$$

$$= 5.96$$

$$WGSS_2 = (2\text{-}5)^2 + (1.5\text{-}6)^2 + (2\text{-}5)^2 + (3.5\text{-}6)^2 + (3\text{-}5)^2 + (2.5\text{-}6)^2 + (5\text{-}5)^2 + (6\text{-}6)^2$$

$$= 60.75$$

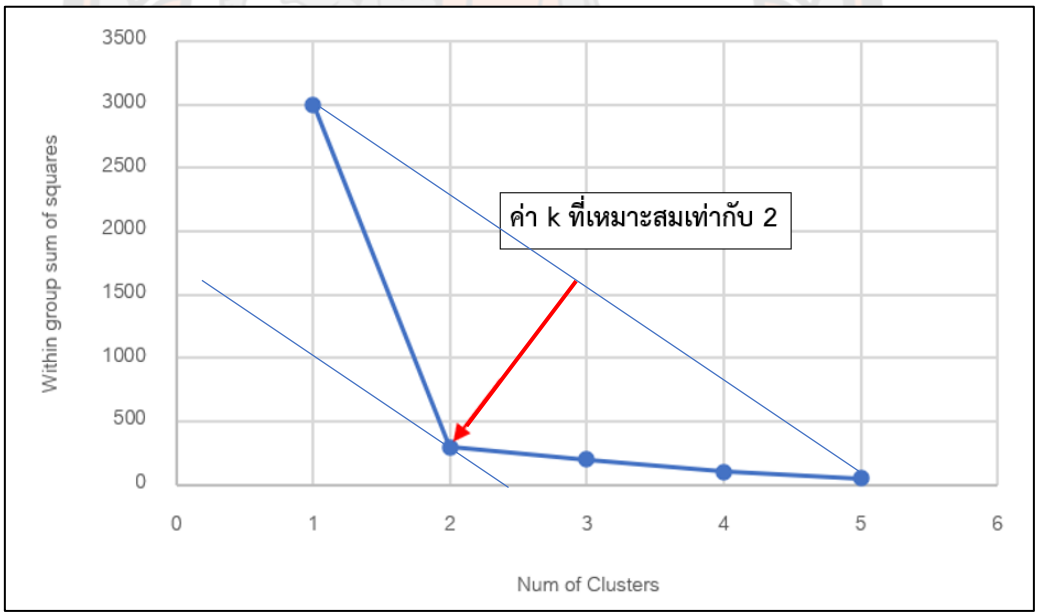
ดังนั้น $WGSS = WGSS_1 + WGSS_2$

$$= 5.96 + 60.75 = 66.71$$

2.3.1 การเลือกค่า k ที่เหมาะสม

การเลือกจำนวนกลุ่มที่เหมาะสมเป็นปัญหาพื้นฐานของเทคนิคการแบ่งกลุ่มข้อมูลแบบเคมีน ที่ไม่มีวิธีการเฉพาะเจาะจงในการกำหนดจำนวนกลุ่มที่แน่นอน โดยจากการทบทวนวรรณพบว่าวิธีที่ใช้ระบุจำนวนกลุ่มที่เหมาะสมที่ได้รับความนิยมคือวิธี Elbow Method (Aziz Mohammad Nasrul & Ahmad Tohari, 2021)

วิธี Elbow Method เป็นวิธีที่ใช้วัดข้อผิดพลาด (Error Measurement) ของผลรวมระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางของกลุ่มหรือเรียกว่า WGSS ซึ่งการวนซ้ำแต่ละรอบจะทำให้ค่าของ WGSS ลดลงจากจำนวนคลัสเตอร์ที่เพิ่มขึ้นส่งผลให้จำนวนสมาชิกในแต่ละคลัสเตอร์จะลดลง และเมื่อการคำนวณมีความผิดพลาดน้อยลง ความชันของเส้นโค้งจะเริ่มเรียบ (Smooth) และจะเกิดเป็นมุมที่มีลักษณะคล้ายข้อศอก (Elbow) โดยการพิจารณาหาจำนวนคลัสเตอร์ที่เหมาะสมที่สุดนั้นให้ลากเส้นจากจุดเริ่มไปยังปลายเส้นโค้ง จากนั้นหาระยะจากเส้นตรงตั้งฉากกับเส้นโค้งที่มีระยะห่างมากที่สุดก็จะได้จำนวนคลัสเตอร์ที่เหมาะสมที่สุด ดังภาพที่ 16 จะเห็นได้ว่าจำนวนคลัสเตอร์เท่ากับ 2 เป็นค่าที่เหมาะสมที่สุด (สุกฤษฎิ์ ไกรนรา, 2563) (Cui, 2020)



ภาพที่ 16 การเลือกจำนวนกลุ่มด้วยวิธี Elbow Method

2.4 งานวิจัยที่เกี่ยวข้อง

2.4.1 ด้านเทคนิคการจำแนก

Comert and Kocamaz (2017) ศึกษาและเปรียบเทียบประสิทธิภาพการจำแนกอัตราการเต้นหัวใจของทารกในครรภ์ด้วยเทคนิคการเรียนรู้ของเครื่อง สำหรับข้อมูลที่ใช้ คือ ข้อมูลอัตราการเต้นหัวใจของทารกในครรภ์จำนวน 2,126 ชุด โดยแบ่งออกเป็น 2 กลุ่มได้แก่ กลุ่มทารกที่มีอัตราการเต้นหัวใจแบบปรกติและทารกที่มีอัตราการเต้นหัวใจผิดปรกติ (ภาวะขาดออกซิเจน) โดยเทคนิคการจำแนกที่ใช้ในงานวิจัยประกอบด้วย โครงข่ายประสาทเทียม เครื่องสนับสนุนเวกเตอร์ เครื่องจักรเรียนรู้เอ็กซ์ทรีม และเทคนิคป่าสุ่ม ผลจากงานวิจัยพบว่า เทคนิคโครงข่ายประสาทเทียมมีประสิทธิภาพในการจำแนกสูงที่สุด โดยมีค่าความไว (Sensitivity) ร้อยละ 99.73 ค่าความจำเพาะ (Specificity) ร้อยละ 97.94 และค่าวัดประสิทธิภาพ 99.70

Kublanov et al. (2017) ศึกษาและเปรียบเทียบประสิทธิภาพการจำแนกผู้ป่วยโรคความดันโลหิตสูงด้วยเทคนิคการเรียนรู้ของเครื่อง สำหรับข้อมูลที่ใช้คือ ชุดข้อมูลผู้ป่วยจำนวน 70 ชุด โดยแบ่งออกเป็น 2 กลุ่มได้แก่ กลุ่มผู้ป่วยโรคความดันโลหิตสูงและกลุ่มผู้ที่ไม่ได้ป่วยโรคความดันโลหิตสูง โดยเทคนิคการจำแนกที่ใช้ในงานวิจัยประกอบด้วย การวิเคราะห์จำแนกกลุ่มเชิงเส้น การวิเคราะห์จำแนกกลุ่มกำลังสอง (Quadratic Discriminant Analysis) เพื่อนบ้านใกล้ที่สุด เครื่องสนับสนุนเวกเตอร์ ต้นไม้และเทคนิคนาอีฟเบย์ ผลจากงานวิจัยพบว่า การวิเคราะห์จำแนกกลุ่มเชิงเส้นมีประสิทธิภาพในการจำแนกสูงที่สุด โดยมีค่าความแม่นร้อยละ 91.3

Hutapea et al. (2019) (Hutapea et al., 2019) ศึกษาและเปรียบเทียบประสิทธิภาพการจำแนกทิศทางของราคาหุ้น Pt Astra International tkb ด้วยเทคนิคการเรียนรู้ของเครื่อง สำหรับข้อมูลที่ใช้ คือ ชุดข้อมูลราคาหุ้นของ Pt Astra International tkb จากตลาดหลักทรัพย์อินโดนีเซียจำนวน 1,195 ชุด โดยแบ่งออกเป็น 2 กลุ่ม คือ กลุ่มราคาหุ้นขึ้นและกลุ่มราคาหุ้นลง โดยเทคนิคการจำแนกที่ใช้ในงานวิจัยประกอบด้วย เทคนิคนาอีฟเบย์ และเทคนิคต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 ผลจากงานวิจัยพบว่า เทคนิคต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 มีประสิทธิภาพในการจำแนกสูงที่สุด โดยมีค่าความแม่นร้อยละ 98.74 ค่าความเที่ยงร้อยละ 98.90 และค่าการเรียกคืนร้อยละ 99.70

Cinar and Koklu (2019) (Cinar & Koklu, 2019) ศึกษาและเปรียบเทียบประสิทธิภาพการจำแนกสายพันธุ์ข้าวด้วยเทคนิคการเรียนรู้ของเครื่อง สำหรับข้อมูลที่ใช้คือ ชุดข้อมูลสายพันธุ์ข้าวจำนวน 3,810 ชุด โดยแบ่งออกเป็น 2 กลุ่ม คือ กลุ่มข้าวสายพันธุ์ Osmancik และกลุ่มข้าวสายพันธุ์ Cammeo โดยเทคนิคการจำแนกที่ใช้ในงานวิจัยประกอบด้วย เทคนิคการถดถอยลอจิสติก โครงข่ายประสาทเทียมแบบหลายชั้น เวกเตอร์ค้ำยัน ต้นไม้ตัดสินใจ เทคนิคป่าสุ่ม เทคนิคนาอีฟเบย์ และเพื่อนบ้านใกล้ที่สุด ผล

จากงานวิจัยพบว่า เทคนิคการถดถอยลอจิสติก มีประสิทธิในการจำแนกสูงที่สุด โดยมีค่าความแม่นร้อยละ 93.02 ค่าการเรียกคืนร้อยละ 92.26 ค่าความจำเพาะร้อยละ 93.58 ค่าความเที่ยงร้อยละ 91.35

Religia et al. (2020) (Religia et al., 2020) ศึกษาและสร้างตัวแบบการจำแนกเพื่อประเมินความเสี่ยงในการให้เครดิตด้วยเทคนิคการเรียนรู้ของเครื่อง สำหรับข้อมูลที่ใช้คือ ชุดข้อมูลลูกค้าของสถาบันการเงินในประเทศเยอรมันจำนวน 45,211 ชุด โดยแบ่งออกเป็น 2 กลุ่มคือ กลุ่มลูกค้าที่ปฏิบัติตามสัญญาสินเชื่อและกลุ่มลูกค้าที่ไม่ปฏิบัติตามสัญญาสินเชื่อ โดยเทคนิคการจำแนกที่ใช้ในงานวิจัยคือ เทคนิคป่าสุ่ม ผลจากงานวิจัยพบว่า เทคนิคป่าสุ่มให้ค่าความแม่นร้อยละ 78.33

Golpour et al. (2020) ศึกษาและเปรียบเทียบประสิทธิภาพการจำแนกผู้ป่วยโรคหลอดเลือดหัวใจตีบด้วยเทคนิคการเรียนรู้ของเครื่อง สำหรับข้อมูลคือ ชุดข้อมูลผู้ป่วยจำนวน 1,141 ชุด โดยแบ่งออกเป็น 2 กลุ่มคือ กลุ่มผู้ที่ป่วยและกลุ่มผู้ที่ไม่ได้ป่วยเป็นโรคหลอดเลือดหัวใจตีบ โดยเทคนิคการจำแนกที่ใช้ในงานวิจัยประกอบด้วย เครื่องสนับสนุนเวกเตอร์ เทคนิคนาอีฟเบย์ และการถดถอยลอจิสติก ผลจากงานวิจัยพบว่า เทคนิคนาอีฟเบย์มีประสิทธิภาพในการจำแนกสูงที่สุด โดยมีค่าการเรียกคืนร้อยละ 89.20 ค่าความจำเพาะร้อยละ 42.80 ค่า AUC ร้อยละ 74.00 และค่าความแม่นร้อยละ 71.30

Tarakci and Ozkan (2021) (Tarakci & Ozkan, 2021) ศึกษาและเปรียบเทียบประสิทธิภาพการจำแนกสายพันธุ์ข้าวด้วยเทคนิคการเรียนรู้ของเครื่อง สำหรับข้อมูลที่ใช้คือ ชุดข้อมูลลักษณะข้าวจำนวน 3,810 ชุด โดยแบ่งออกเป็น 2 กลุ่ม คือ กลุ่มข้าวสายพันธุ์ Osmancik และกลุ่มข้าวสายพันธุ์ Cammeo โดยเทคนิคการจำแนกที่ใช้ในงานวิจัยประกอบด้วย เทคนิคเพื่อนบ้านใกล้ที่สุดและเพื่อนบ้านใกล้ที่สุดแบบถ่วงน้ำหนัก (Weighted K-Nearest Neighbors) ผลจากงานวิจัยพบว่า เทคนิคเพื่อนบ้านใกล้ที่สุดมีประสิทธิภาพในการจำแนกสูงที่สุด โดยมีค่าความแม่นร้อยละ 92.60 ค่า Error rate ร้อยละ 7.40 ค่าความจำเพาะร้อยละ 94.77 ค่าความแม่นร้อยละ 92.76 ค่าการเรียกคืนร้อยละ 89.63 และค่าประสิทธิภาพร้อยละ 91.17

2.4.2 ด้านการจัดการปัญหาข้อมูลอสมดุล

กาญจน์ ณ ศรีธะ (2560) ศึกษาและเปรียบเทียบประสิทธิภาพการจำแนกด้วยเทคนิคการเรียนรู้ของเครื่อง สำหรับข้อมูลที่ใช้คือ ชุดข้อมูลผู้ป่วยโรคมะเร็งเต้านม (Breast Cancer) ชุดข้อมูลผู้ป่วยโรคเบาหวานประเภทที่ 2 (Pima Indian Diabetes) และชุดข้อมูลผู้ป่วยโรคเท้าเบาหวาน (Diabetic foot) โดยเทคนิคที่ใช้ปรับปรุงชุดข้อมูลอสมดุลให้สมดุลประกอบด้วย วิธีสังเคราะห์ข้อมูลใหม่ (Synthetic Minority Over-Sampling Technique) วิธีสุ่มลด (Under Sampling) และเทคนิคการสุ่มตัวอย่างซ้ำ เทคนิคการจำแนกที่ใช้ประกอบด้วย ต้นไม้ตัดสินใจ คาร์ท (CART) เทคนิคป่าสุ่ม เครื่องสนับสนุนเวกเตอร์และโครงข่ายประสาทเทียม และเทคนิคที่ใช้ช่วยเพิ่มประสิทธิภาพการจำแนกได้แก่ เทคนิคเอดาบูทและเทคนิคถุงจำแนก ผลจากงานวิจัยพบว่า เมื่อพิจารณาที่ชุดข้อมูล

ผู้ป่วยโรคมะเร็งเต้านม เทคนิคป่าสุ่มโดยปรับปรุงชุดข้อมูลอสมดุลให้สมดุลด้วยเทคนิคการสุ่มตัวอย่างซ้ำมีประสิทธิภาพในการจำแนกสูงที่สุด โดยมีค่าความเที่ยงร้อยละ 86.38 ค่าการเรียกคืนร้อยละ 86.64 ค่าประสิทธิภาพร้อยละ 86.40 เมื่อพิจารณาที่ชุดข้อมูลผู้ป่วยโรคเบาหวานประเภทที่ 2 เทคนิคป่าสุ่มโดยปรับปรุงข้อมูลอสมดุลให้สมดุลด้วยเทคนิคการสุ่มตัวอย่างซ้ำมีประสิทธิภาพในการจำแนกสูงที่สุด โดยมีค่าความเที่ยงร้อยละ 91.25 ค่าการเรียกคืนร้อยละ 91.28 ค่าประสิทธิภาพร้อยละ 91.19 และเมื่อพิจารณาที่ชุดข้อมูลผู้ป่วยโรคเท้าเบาหวาน เทคนิคป่าสุ่มที่ทำงานร่วมกับเทคนิคเอดาบูทมีประสิทธิภาพในการจำแนกสูงที่สุด โดยมีค่าความเที่ยงร้อยละ 86.46 ค่าการเรียกคืนร้อยละ 86.41 ค่าประสิทธิภาพร้อยละ 85.91

Verma (2019) (Verma, 2019) ศึกษาการปรับปรุงชุดข้อมูลอสมดุลให้สมดุลและเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มลูกค้าของสถาบันการเงินด้วยเทคนิคการเรียนรู้ของเครื่อง สำหรับข้อมูลที่ใช้คือ ชุดข้อมูลลูกค้าของสถาบันการเงินในโปตุเกสจำนวน 45,211 ชุด โดยแบ่งออกเป็น 2 กลุ่มได้แก่ กลุ่มลูกค้าที่สมัครเงินฝากประจำและกลุ่มลูกค้าที่ไม่สมัครเงินฝากประจำ โดยเทคนิคที่ใช้ปรับปรุงชุดข้อมูลอสมดุลให้สมดุลประกอบด้วย วิธีสุ่มลด วิธีสุ่มเกิน และวิธีสังเคราะห์ข้อมูลใหม่ เทคนิคการจำแนกที่ใช้ประกอบด้วย ต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 เทคนิคนาอีฟเบย์ เครื่องสนับสนุนเวกเตอร์ การถดถอยลอจิสติก โครงข่ายประสาทเทียมแบบหลายชั้น และเทคนิคป่าสุ่ม ผลจากงานวิจัยพบว่า เทคนิคป่าสุ่มโดยปรับข้อมูลอสมดุลให้สมดุลด้วยวิธีสุ่มเกินมีประสิทธิภาพการจำแนกสูงที่สุด โดยมีค่าความเที่ยงร้อยละ 92.20 ค่าความการเรียกคืนร้อยละ 98.70 ค่าประสิทธิภาพร้อยละ 95.30 พื้นที่ใต้เส้นโค้ง ROC ร้อยละ 99.40

Aziz Mohammad and Ahmad Tohari (2021) (Aziz Mohammad Nasrul & Ahmad Tohari, 2021) ศึกษาปรับปรุงชุดข้อมูลอสมดุลให้สมดุลและเปรียบเทียบประสิทธิภาพการจำแนกด้วยเทคนิคการเรียนรู้ของเครื่อง สำหรับข้อมูลที่ใช้คือ ชุดข้อมูล NSL-KDD และชุดข้อมูล UNSW-NB15 โดยเทคนิคที่ใช้ปรับปรุงชุดข้อมูลอสมดุลให้สมดุลประกอบด้วย การประยุกต์ใช้วิธีการแบ่งกลุ่มข้อมูลแบบเคมีน เทคนิคการจำแนกที่ใช้ประกอบด้วย เทคนิคนาอีฟเบย์ เพื่อนบ้านใกล้ที่สุด เครื่องสนับสนุนเวกเตอร์ เพอร์เซ็ปตรอนหลายชั้น (Multi-Layer Perceptron) และเทคนิคป่าสุ่ม ผลจากงานวิจัยพบว่า เมื่อพิจารณาชุดข้อมูล NSL-KDD เทคนิคป่าสุ่มมีประสิทธิภาพในการจำแนกสูงที่สุด โดยมีค่าความแม่นร้อยละ 99.76 ค่าความเที่ยงร้อยละ 99.80 ค่าการเรียกคืนร้อยละ 99.80 ตามลำดับ และเมื่อพิจารณาชุดข้อมูล UNSW-NB15 เทคนิคเพื่อนบ้านใกล้ที่สุดมีประสิทธิภาพในการจำแนกสูงที่สุด โดยมีค่าความแม่นร้อยละ 98.96 ค่าความเที่ยงร้อยละ 99.00 ค่าการเรียกคืนร้อยละ 99.00

พัชรียา ทองพูล (2562) ศึกษาการปรับปรุงชุดข้อมูลอสมดุลให้สมดุลและเปรียบเทียบประสิทธิภาพการจำแนกด้วยเทคนิคการเรียนรู้ของเครื่อง สำหรับชุดข้อมูลที่ใช้คือ ชุดข้อมูลการรับรู้

ทางหูของเด็ก ชุดข้อมูลยอดคงเหลือในบัตรเครดิตของลูกค้า และชุดข้อมูลคุณภาพไวน์แดง โดยมีจำนวน 890 400 999 ชุด ตามลำดับ โดยเทคนิคที่ปรับปรุงชุดข้อมูลอสมดุลให้สมดุลประกอบด้วย วิธีการสุ่มเกิน (Over Sampling) วิธีสุ่มเกินโดยใช้เทคนิค SMOTE (Synthetic Minority Over-Sampling) วิธีสังเคราะห์ข้อมูลใหม่ วิธีสุ่มลด และวิธีผสมผสาน เทคนิคการจำแนกที่ใช้ประกอบด้วย เพื่อนบ้านใกล้ที่สุด ต้นไม้ตัดสินใจ โครงข่ายประสาทเทียม และเครื่องสนับสนุนเวกเตอร์ ผลจากงานวิจัยพบว่า เมื่อพิจารณาชุดข้อมูลการรับรู้ทางหูของเด็กพบว่า เครื่องสนับสนุนเวกเตอร์โดยการแก้ไขข้อมูลอสมดุลด้วยเทคนิค SMOTE มีประสิทธิภาพในการจำแนกสูงที่สุด โดยมีค่าความแม่นร้อยละ 85.09 ค่ากาเรียกคืนร้อยละ 82.10 ค่าความเที่ยงร้อยละ 88.38 และค่าความคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 0.1490 ตามลำดับ เมื่อพิจารณาชุดข้อมูลยอดคงเหลือในบัตรเครดิตของลูกค้าพบว่า เทคนิคเพื่อนบ้านใกล้ที่สุดโดยการแก้ไขข้อมูลอสมดุลด้วยเทคนิค SMOTE มีประสิทธิภาพในการจำแนกสูงที่สุด โดยมีค่าความแม่นร้อยละ 77.78 ค่าการเรียกคืนร้อยละ 78.34 ค่าความเที่ยงร้อยละ 76.80 และค่าความคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 0.2157 ตามลำดับ เมื่อพิจารณาชุดข้อมูลคุณภาพไวน์แดงพบว่า เครื่องสนับสนุนเวกเตอร์โดยการปรับปรุงชุดข้อมูลอสมดุลให้สมดุลด้วยวิธีสุ่มลดมีประสิทธิภาพในการจำแนกสูงที่สุด โดยมีค่าความแม่นร้อยละ 65.13 ค่าการเรียกคืนร้อยละ 65.00 ค่าความเที่ยงร้อยละ 0.5293 และค่าความคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 0.3487

วิชญ์วิสิฐ เกสรสิทธิ์ (2561) ศึกษาปรับปรุงชุดข้อมูลอสมดุลให้สมดุลและเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มการกลับมารักษาซ้ำของผู้ป่วยโรคเบาหวานด้วยเทคนิคการเรียนรู้ของเครื่อง สำหรับชุดข้อมูลที่ใช้คือ การกลับมารักษาซ้ำของผู้ป่วยโรคเบาหวานจำนวน 101,766 ชุด โดยแบ่งออกเป็น 3 กลุ่มได้แก่ กลุ่มผู้ป่วยที่ไม่กลับมารักษาซ้ำ กลุ่มผู้ป่วยที่กลับมารักษาซ้ำภายใน 30 วัน และกลุ่มผู้ป่วยที่กลับมารักษามากกว่า 30 วัน โดยเทคนิคที่ใช้แก้ปัญหาอสมดุลคือ วิธีสุ่มเกิน วิธีสุ่มลด วิธีผสมผสาน และวิธีสังเคราะห์ข้อมูลใหม่ เทคนิคการจำแนกที่ใช้ประกอบด้วย โดยใช้การถดถอยลอจิสติกและต้นไม้ตัดสินใจ ผลจากงานวิจัยพบว่า ต้นไม้ตัดสินใจโดยการปรับปรุงชุดข้อมูลอสมดุลให้สมดุลด้วยวิธีสังเคราะห์ข้อมูลใหม่มีประสิทธิภาพในการจำแนกสูงที่สุด

Morteza Pakdaman (2020) ศึกษาการปรับปรุงชุดข้อมูลอข้อมูลอสมดุลให้สมดุลและเปรียบเทียบประสิทธิภาพการจำแนกด้วยเทคนิคการเรียนรู้ของเครื่อง สำหรับข้อมูลที่ใช้คือ ชุดข้อมูลการเกิดฟ้าผ่าที่เมือง Mashhad ชุดข้อมูลการเกิดฟ้าผ่าที่เมือง Neyshabour และชุดข้อมูลการเกิดฟ้าผ่าที่เหมือง Quchan โดยเทคนิคที่ใช้ปรับปรุงชุดข้อมูลอสมดุลให้สมดุลคือ วิธีสุ่มลด เทคนิคการจำแนกที่ใช้ประกอบด้วย โครงข่ายประสาทเทียมและต้นไม้ตัดสินใจ ผลจากงานวิจัยพบว่า เมื่อพิจารณาที่ชุดข้อมูลการเกิดฟ้าผ่าที่เมือง Mashhad ต้นไม้ตัดสินใจมีประสิทธิภาพในการจำแนกมากที่สุด โดยมีค่าความแม่นร้อยละ 86.70 ค่าความเที่ยงร้อยละ 86.50 ค่าการเรียกคืนร้อยละ 87.20 ค่าประสิทธิภาพร้อยละ 86.80 เมื่อพิจารณาที่ชุดข้อมูลการเกิดฟ้าผ่าที่เมือง Neyshabour ต้นไม้

ตัดสินใจมีประสิทธิภาพในการจำแนกมากที่สุด โดยมีค่าความแม่นร้อยละ 85.70 ค่าความเที่ยงร้อยละ 86.20 ค่าการเรียกคืนร้อยละ 85.00 ค่าประสิทธิภาพร้อยละ 85.60 เมื่อพิจารณาที่ชุดข้อมูลการเกิดฟ้าผ่าที่เหมือง Quchan ต้นไม้ตัดสินใจมีประสิทธิภาพในการจำแนกมากที่สุด โดยมีค่าความแม่นร้อยละ 85.60 ค่าความเที่ยงร้อยละ 85.40 ค่าการเรียกคืนร้อยละ 85.80 ค่าประสิทธิภาพร้อยละ 85.60

Fayaz Itoo (2021) ศึกษาการปรับปรุงชุดข้อมูลอสมดุลให้สมดุลและเปรียบเทียบประสิทธิภาพการจำแนกการฉ้อโกงบัตรเครดิตด้วยเทคนิคการเรียนรู้ของเครื่อง สำหรับข้อมูลที่ใช้คือข้อมูลตรวจจับการฉ้อโกงบัตรเครดิตจำนวน 284,807 ชุด โดยแบ่งออกเป็น 2 กลุ่มได้แก่ กลุ่มที่ฉ้อโกงบัตรเครดิตและกลุ่มที่ไม่ได้ฉ้อโกงบัตรเครดิต โดยเทคนิคที่ใช้ปรับปรุงชุดข้อมูลอสมดุลให้สมดุลคือ วิธีสุ่มลด เทคนิคการจำแนกที่ใช้ประกอบด้วย การถดถอยลอจิสติก เทคนิคนาอีฟเบย์ และเพื่อนบ้านใกล้ที่สุด ผลจากงานวิจัยพบว่า การถดถอยลอจิสติกมีประสิทธิภาพในการจำแนกสูงที่สุด โดยมีค่าความแม่นร้อยละ 95.90 ค่าการเรียกคืนร้อยละ 83.90 ค่าความจำเพาะร้อยละ 99.70 ค่าความเที่ยงร้อยละ 99.10 ค่าประสิทธิภาพร้อยละ 90.90 ค่า AUC ร้อยละ 91.80

ตารางที่ 5 แสดงผลสรุปงานวิจัยที่เกี่ยวข้องด้านเทคนิคการจำแนกและการจัดการปัญหาข้อมูลอสมดุล

งานวิจัย	ปี	เทคนิคการจำแนก						อสมดุล	
		LDA	NB	DT	RF	ANN	Other	Over	Under
Comert	2017				✓	✓	✓		
Kublanov et al.	2017	✓	✓	✓			✓		
Hutapea et al.	2019		✓	✓					
Cinar and Koklu	2019		✓	✓	✓	✓	✓		
Religia et al.	2020				✓				
Golpour et al.	2020		✓				✓		
Tarakci and Ozkan	2021						✓		
กาญจน์ ณ ศรีธะ	2560			✓	✓	✓	✓	✓	✓
Verma	2019		✓	✓	✓	✓	✓	✓	✓
Mohammad	2021		✓		✓		✓		✓
พัชรียา ทองพูล	2562			✓		✓	✓	✓	✓
วิชญ์วิสิฐ เกสรสิทธิ์	2561			✓			✓	✓	✓
Morteza Pakdaman	2020			✓		✓			✓
Fayaz Itoo	2021		✓				✓		✓

จากตารางที่ 5 แสดงผลสรุปงานวิจัยที่เกี่ยวข้อง โดยกำหนดเครื่องหมายถูก หมายถึงเทคนิคที่ใช้ในงานวิจัยและเครื่องหมายถูกสีแดง หมายถึงเทคนิคที่มีประสิทธิภาพสูงที่สุด

บทที่ 3
วิธีการดำเนินการวิจัย

3.1 ข้อมูลที่ใช้ในการศึกษา

ในงานวิจัยนี้ได้คัดเลือกชุดข้อมูลจาก UCI Machine Learning Repository โดยข้อมูลแต่ละชุดมีลักษณะดังต่อไปนี้

1. ชุดข้อมูลที่ 1 คือ ข้อมูลสถาบันการเงินมีจำนวนข้อมูลทั้งหมด 30,488 ชุด ภายในชุดข้อมูลประกอบด้วยตัวแปรอิสระ 20 ตัว ซึ่งเป็นตัวแปรอิสระเชิงคุณภาพ 10 ตัว และตัวแปรอิสระเชิงปริมาณ 10 ตัว และมีตัวแปรตามเชิงกลุ่ม 1 ตัว ซึ่งแบ่งเป็น 2 กลุ่มได้แก่ กลุ่มลูกค้าที่ต้องการสมัครบัญชีเงินฝากประจำจำนวน 3,859 ชุด (โดยในที่นี้จะกำหนดให้เป็นกลุ่ม Yes) และกลุ่มลูกค้าที่ไม่ต้องการสมัครบัญชีเงินฝากประจำจำนวน 26,629 ชุด (โดยในที่นี้จะกำหนดให้เป็นกลุ่ม No)

2. ชุดข้อมูลที่ 2 คือ ข้อมูลสายพันธุ์ข้าวมีจำนวนข้อมูลทั้งหมด 3,810 ชุด ภายในชุดข้อมูลประกอบด้วยตัวแปรอิสระ 7 ตัว ซึ่งเป็นตัวแปรอิสระเชิงคุณภาพ 0 ตัว และตัวแปรอิสระเชิงปริมาณ 7 ตัว และมีตัวแปรตามเชิงกลุ่ม 1 ตัว ซึ่งแบ่งเป็น 2 กลุ่มได้แก่ กลุ่มข้าวสายพันธ์ Cammeo จำนวน 1,630 ชุด (โดยในที่นี้จะกำหนดให้เป็นกลุ่ม Yes) และกลุ่มข้าวสายพันธ์ Osmancik จำนวน 2,180 ชุด (โดยในที่นี้จะกำหนดให้เป็นกลุ่ม No)

3. ชุดข้อมูลที่ 3 คือ ข้อมูลนักวิทยาศาสตร์ข้อมูลมีจำนวนข้อมูลทั้งหมด 8,955 ชุด ภายในชุดข้อมูลประกอบด้วยตัวแปรอิสระ 10 ตัว ซึ่งเป็นตัวแปรเชิงคุณภาพ 9 ตัว และตัวแปรเชิงปริมาณ 1 ตัว และมีตัวแปรตามเชิงกลุ่ม 1 ตัว ซึ่งแบ่งเป็น 2 กลุ่มได้แก่ กลุ่มคนที่ต้องการเปลี่ยนอาชีพมาเป็นนักวิทยาศาสตร์ข้อมูลจำนวน 1,483 ชุด (โดยในที่นี้จะกำหนดให้เป็นกลุ่ม Yes) และกลุ่มคนที่ไม่ต้องการเปลี่ยนอาชีพมาเป็นนักวิทยาศาสตร์จำนวน 7,472 ชุด (โดยในที่นี้จะกำหนดให้เป็นกลุ่ม No)

โดยรายละเอียดชุดข้อมูลตั้งต้นของแต่ละชุดข้อมูลสรุปได้ดังตารางที่ 6

ตารางที่ 6 รายละเอียดชุดข้อมูลตั้งต้นที่ใช้ในงานวิจัย

ชื่อชุดข้อมูล	จำนวนตัวแปรอิสระ		ค่าตัวแปรตาม		จำนวนข้อมูล	
	เชิงคุณภาพ	เชิงปริมาณ	Yes	No	จำนวน	สัดส่วนความไม่สมดุล
สถาบันการเงิน	10	10	3,859	26,629	30,488	1:6.7
สายพันธุ์ข้าว	0	7	1,630	2,180	3,810	1:1.3
นักวิทยาศาสตร์ข้อมูล	9	1	1,483	7,472	8,955	1:4.9

3.2 เครื่องมือที่ใช้ในการศึกษา

งานวิจัยนี้ผู้วิจัยเลือกใช้เครื่องมือที่ใช้ในการศึกษาดังนี้

3.2.1 โปรแกรม Microsoft Excel เพื่อใช้ในการคัดกรองข้อมูลเบื้องต้น

3.2.2 โปรแกรม RStudio Version 1.4.1717 เพื่อใช้ในการสร้างตัวแบบการจำแนก

1. แพ็คเกจ MASS เพื่อใช้ในการสร้างตัวแบบเทคนิคการวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์

2. แพ็คเกจ naivebayer เพื่อใช้ในการสร้างตัวแบบเทคนิคนาอีฟเบย์

3. แพ็คเกจ RWeka และ rJava เพื่อใช้ในการสร้างตัวแบบต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5

4. แพ็คเกจ randomForest เพื่อใช้ในการสร้างตัวแบบเทคนิคป่าสุ่ม

5. แพ็คเกจ ANN2 เพื่อใช้ในการสร้างตัวแบบเทคนิคโครงข่ายประสาทเทียม

3.2.3 โปรแกรม Weka (Waikato Environment for Knowledge Analysis) Version 3.8.5 เพื่อใช้ในการวิเคราะห์แบ่งกลุ่มข้อมูลแบบเคมีน

3.3 วิธีวิเคราะห์และจัดเตรียมข้อมูล

การวิจัยนี้มีจุดประสงค์เพื่อเปรียบเทียบตัวแบบการจำแนกโดยใช้เทคนิคทางสถิติและเทคนิคการเรียนรู้ของเครื่องประกอบด้วย เทคนิคการวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์ เทคนิคนาอีฟเบย์ เทคนิคต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 เทคนิคป่าสุ่มและเทคนิคโครงข่ายประสาทเทียม ภายใต้ชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพและเชิงปริมาณที่แตกต่างกัน โดยงานวิจัยนี้นำเสนอการปรับปรุงชุดข้อมูลอสมดุลให้สมดุลโดยใช้วิธีสุ่มลด 2 เทคนิคได้แก่ การสุ่มตัวอย่างแบบง่ายและการแบ่งกลุ่มข้อมูลแบบเคมีน โดยมีรายละเอียดการเตรียมข้อมูลแต่ละชุดดังต่อไปนี้

3.3.1 ชุดข้อมูลสถาบันการเงิน

3.3.1.1 ศึกษารายละเอียดชุดข้อมูล

ชุดข้อมูลสถาบันการเงินที่จัดเก็บโดยสถาบันการเงินในประเทศโปรตุเกส ซึ่งจัดทำแคมเปญเพื่อหาลูกค้าที่สนใจสมัครเงินฝากประจำ โดยมีรายละเอียดตัวแปรดังตารางที่ 7

ตารางที่ 7 แสดงรายละเอียดชุดข้อมูลสถาบันการเงิน

ตัวแปร	รายละเอียด
Job (x_1)	อาชีพ (ตัวแปรเชิงคุณภาพ มาตรนามบัญญัติ)
Marital (x_2)	สถานภาพสมรส (ตัวแปรเชิงคุณภาพ มาตรนามบัญญัติ)

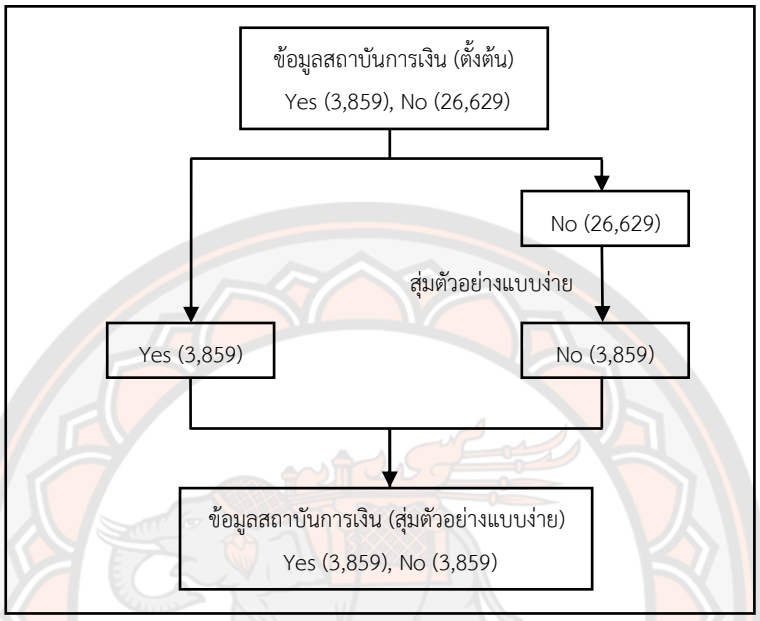
ตัวแปร	รายละเอียด
Education (x_3)	ระดับการศึกษา (ตัวแปรเชิงคุณภาพ มาตรเรียงลำดับ)
Default (x_4)	มีเครดิตผิดนัด (ตัวแปรเชิงคุณภาพ มาตรนามบัญญัติ)
Housing (x_5)	การมีบ้าน (ตัวแปรเชิงคุณภาพ มาตรนามบัญญัติ)
Loan (x_6)	มีสินเชื่อส่วนบุคคลหรือไม่ (ตัวแปรเชิงคุณภาพ มาตรนามบัญญัติ)
Contact (x_7)	การติดต่อ (ตัวแปรเชิงคุณภาพ มาตรนามบัญญัติ)
Month (x_8)	เดือน (ตัวแปรเชิงคุณภาพ มาตรนามบัญญัติ)
Day_of_week (x_9)	วันติดต่อ (ตัวแปรเชิงคุณภาพ มาตรนามบัญญัติ)
Poutcome (x_{10})	ผลลัพธ์ของแคมเปญก่อนหน้า (ตัวแปรเชิงคุณภาพ มาตรเรียงลำดับ)
Duration (x_{11})	ระยะเวลาที่ใช้พูดคุยกับลูกค้า (ตัวแปรเชิงปริมาณ)
Age (x_{12})	อายุ (ตัวแปรเชิงปริมาณ)
Campaign (x_{13})	จำนวนผู้ติดต่อที่ดำเนินการสำหรับลูกค้ารายนี้ (ตัวแปรเชิงปริมาณ)
Pdays (x_{14})	จำนวนวันหลังจากที่ลูกค้าได้รับการติดต่อครั้งสุดท้ายจากแคมเปญก่อนหน้า (ตัวแปรเชิงปริมาณ)
Previous (x_{15})	จำนวนผู้ติดต่อที่ดำเนินการสำหรับลูกค้ารายนี้ ในแคมเปญก่อนหน้า (ตัวแปรเชิงปริมาณ)
Emp.var.rate (x_{16})	ดัชนีความเชื่อมั่นผู้บริโภค – ตัวบ่งชี้รายเดือน (ตัวแปรเชิงปริมาณ)
Cons.price.idx (x_{17})	ดัชนีราคาผู้บริโภค – ตัวบ่งชี้รายเดือน (ตัวแปรเชิงปริมาณ)
Cons.conf.idx (x_{18})	ดัชนีความเชื่อมั่นผู้บริโภค – ตัวบ่งชี้รายเดือน (ตัวแปรเชิงปริมาณ)
Euribor3m (x_{19})	อัตรา euribor 3 เดือน – ตัวบ่งชี้รายวัน (ตัวแปรเชิงปริมาณ)
Nr.employed (x_{20})	จำนวนพนักงาน (ตัวแปรเชิงปริมาณ)
Y	ลูกค้าสมัครเงินฝากประจำหรือไม่ (ตัวแปรเชิงคุณภาพ)

เริ่มต้นชุดข้อมูลสถาบันการเงินมีจำนวนทั้งหมด 41,188 ชุด ซึ่งภายในชุดข้อมูลจะมีบางข้อมูลที่เป็นค่าสูญหาย (Unknow) ผู้วิจัยได้ทำการลบข้อมูลดังกล่าวออกทำให้เหลือข้อมูลจำนวน 30,488 ชุด

3.3.1.2 การปรับปรุงชุดข้อมูลอสมดุลให้สมดุล

ในงานวิจัยนี้ผู้วิจัยจะใช้วิธีสุ่มลด (Under-Sampling) โดยใช้การสุ่มตัวอย่างแบบง่าย (Simple Random Sampling) และการแบ่งกลุ่มข้อมูลแบบเคมีน (k-means) เพื่อปรับปรุงชุดข้อมูลอสมดุลให้สมดุล โดยมีรายละเอียดดังต่อไปนี้

1. การสุ่มตัวอย่างแบบง่าย คือการสุ่มตัวอย่างโดยกำหนดให้ทุกหน่วยมีโอกาสถูกเลือกเท่า ๆ กันแบบไม่แทนที่ โดยในที่นี้ทำการสุ่มลดจำนวนข้อมูลของกลุ่ม No ให้มีขนาดเท่ากับกลุ่ม Yes แสดงดังในภาพที่ 17



ภาพที่ 17 การปรับปรุงชุดข้อมูลอสมดุลให้สมดุลด้วยวิธีการสุ่มตัวอย่างแบบง่ายของชุดข้อมูลสถาบันการเงิน

2. การแบ่งกลุ่มข้อมูลแบบเคมีน โดยจะพิจารณาลดจำนวนของกลุ่ม No โดยมีขั้นตอนดังต่อไปนี้

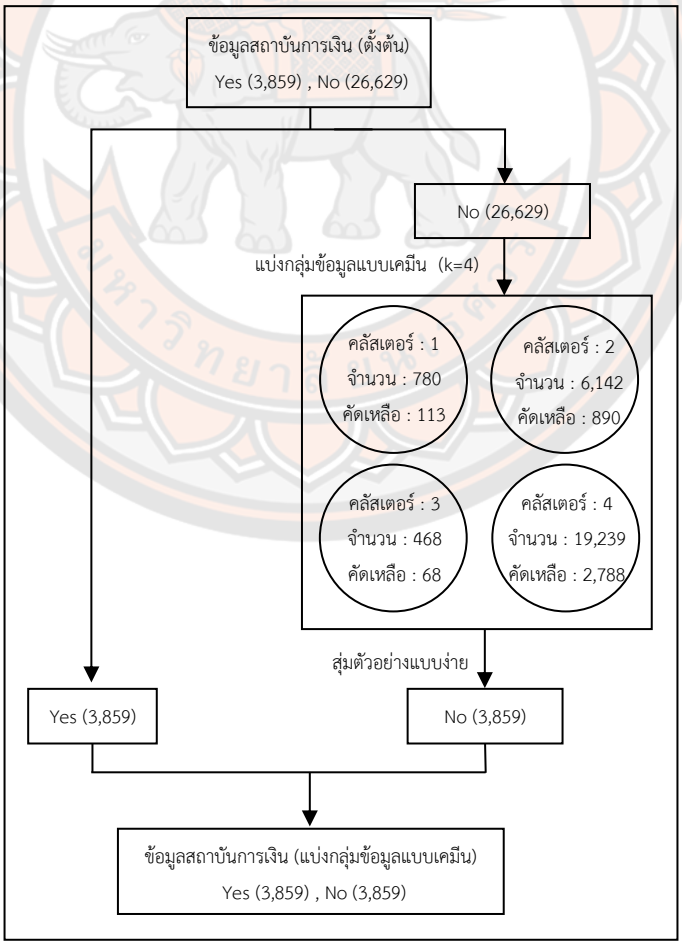
1) ดำเนินการแบ่งกลุ่มด้วยเทคนิคการแบ่งกลุ่มข้อมูลแบบเคมีนเป็น k กลุ่มโดยพิจารณาหาค่า k ที่เหมาะสมด้วยวิธี Elbow Method

2) สุ่มเลือกสมาชิกของแต่ละกลุ่ม (k) โดยเป็นการสุ่มตัวอย่างแบบง่ายแบบไม่แทนที่และสุ่มจำนวนข้อมูลตามสัดส่วนของสมาชิกแต่ละกลุ่ม

3) รวบรวมข้อมูลที่ถูกสุ่ม

ภายใต้ชุดข้อมูลสถาบันการเงิน พบว่าค่า k ที่เหมาะสมมีค่าเท่ากับ 4 โดยวิธี Elbow Method ดังภาพที่ 18 จากนั้นทำการสุ่มตัวอย่างแบบง่ายตามสัดส่วนของสมาชิกแต่ละคลัสเตอร์ยกตัวอย่างเช่น ทำการสุ่มตัวอย่างแบบง่ายจากคลัสเตอร์ที่ 1 จำนวน $\left(\frac{780}{26,629}\right)\times 3,859 = 113$ และทำเช่นเดียวกันกับคลัสเตอร์ที่เหลือดังภาพที่ 19

ภาพที่ 18 ภายใต้ชุดข้อมูลสถาบันการเงินค่า k ที่เหมาะสมคือ 4 โดยวิธีของ Elbow Method



ภาพที่ 19 การปรับปรุงชุดข้อมูลอสมดุลให้สมดุลด้วยเทคนิคการแบ่งกลุ่มข้อมูลแบบเคมีน ของชุดข้อมูลสถาบันการเงิน เมื่อค่า k ที่เหมาะสมเท่ากับ 4

3.3.1.3 ทำการแปลงตัวแปรอิสระให้เป็นปรกติและเป็นตัวแปรดัมมี่

ในการพัฒนาตัวแบบการจำแนกในงานวิจัยนี้ผู้วิจัยจะมีการแปลงตัวแปรอิสระ โดยตัวแปรเชิงปริมาณและตัวแปรเชิงคุณภาพมาตรเรียงลำดับ (Ordinal Scale) จะปรับให้เป็นปรกติ มีค่าระหว่าง 0 – 1 และตัวแปรเชิงคุณภาพมาตรนามบัญญัติ (Nominal Scale) จะแปลงให้เป็นตัวแปรดัมมี่แบบไบนารี่ โดยมีรายละเอียดดังตารางต่อไปนี้

ตัวแปรอิสระก่อนการแปลง	ตัวแปรอิสระหลังการแปลง
1. Job (x_1) (admin, blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed)	1. ทำอาชีพ admin ใช่หรือไม่ (x_{11}) (ใช่ = 1, ไม่ใช่ = 0) 2. ทำอาชีพ blue--collar ใช่หรือไม่ (x_{12}) (ใช่ = 1, ไม่ใช่ = 0) 3. ทำอาชีพ entrepreneur ใช่หรือไม่ (x_{13}) (ใช่ = 1, ไม่ใช่ = 0) 4. ทำอาชีพ housemaid ใช่หรือไม่ (x_{14}) (ใช่ = 1, ไม่ใช่ = 0) 5. ทำอาชีพ management ใช่หรือไม่ (x_{15}) (ใช่ = 1, ไม่ใช่ = 0) 6. ทำอาชีพ retired ใช่หรือไม่ (x_{16}) (ใช่ = 1, ไม่ใช่ = 0) 7. ทำอาชีพ self-employed ใช่หรือไม่ (x_{17}) (ใช่ = 1, ไม่ใช่ = 0) 8. ทำอาชีพ services ใช่หรือไม่ (x_{18}) (ใช่ = 1, ไม่ใช่ = 0) 9. ทำอาชีพ student ใช่หรือไม่ (x_{19}) (ใช่ = 1, ไม่ใช่ = 0) 10. ทำอาชีพ technician ใช่หรือไม่ (x_{110}) (ใช่ = 1, ไม่ใช่ = 0) 11. ทำอาชีพ unemployed ใช่หรือไม่ (x_{111}) (ใช่ = 1, ไม่ใช่ = 0)
2. Marital (x_2) (Single, married, divorced)	12. สถานภาพคือ Single ใช่หรือไม่ (x_{21}) (ใช่ = 1, ไม่ใช่ = 0)

ตัวแปรอิสระก่อนการแปลง	ตัวแปรอิสระหลังการแปลง
	13. สถานภาพคือ married ใช่หรือไม่ (x_{22}) (ใช่ = 1, ไม่ใช่ = 0) 14. สถานภาพคือ divorced ใช่หรือไม่ (x_{23}) (ใช่ = 1, ไม่ใช่ = 0)
3. Education (x_3) (illiterate, basic 4y, basic 9y, high school, professional course, university degree)	15. แปลงให้อยู่ในรูปปรกติโดยมีค่า 0 ถึง 1 (x_3) (max = 6, min = 1)
4. Default (x_4) (เคยผิดนัด, ไม่เคยผิดนัด)	16. มีเครดิตผิดนัดหรือไม่ (x_4) (เคยผิดนัด =1, ไม่เคยผิดนัด = 0)
5. Housing (x_5) (มีบ้าน,ไม่มีบ้าน)	17. การมีบ้าน (x_5) (มีบ้าน = 1, ไม่มีบ้าน = 0)
6. Loan (x_6) (มี , ไม่มี)	18. มีสินเชื่อส่วนบุคคลหรือไม่ (x_6) (มี = 1, ไม่มี = 0)
7. Contact (x_7) (telephone, cellular)	19. การติดต่อ (x_7) (ใช้ telephone ในการติดต่อ = 1, ใช้ cellular ในการติดต่อ = 0)
8. Month (x_8) (apr, aug, dec, jul, jun, mar, may, nov, oct, sep)	20. เดือน apr ใช่หรือไม่ (x_{81}) (ใช่ = 1, ไม่ใช่ = 0) 21. เดือน aug ใช่หรือไม่ (x_{82}) (ใช่ = 1, ไม่ใช่ = 0) 22. เดือน dec ใช่หรือไม่ (x_{83}) (ใช่ = 1, ไม่ใช่ = 0) 23. เดือน jul ใช่หรือไม่ (x_{84}) (ใช่ = 1, ไม่ใช่ = 0) 24. เดือน jun ใช่หรือไม่ (x_{85}) (ใช่ = 1, ไม่ใช่ = 0) 25. เดือน mar ใช่หรือไม่ (x_{86})

ตัวแปรอิสระก่อนการแปลง	ตัวแปรอิสระหลังการแปลง
	(ใช่ = 1, ไม่ใช่ = 0) 26. เดือน may ใช่หรือไม่ (x_{87}) (ใช่ = 1, ไม่ใช่ = 0) 27. เดือน nov ใช่หรือไม่ (x_{88}) (ใช่ = 1, ไม่ใช่ = 0) 28. เดือน oct ใช่หรือไม่ (x_{89}) (ใช่ = 1, ไม่ใช่ = 0) 29. เดือน sep ใช่หรือไม่ (x_{810}) (ใช่ = 1, ไม่ใช่ = 0)
9. Day_of_week (x_9) (mon, thu, wed, tue, fri)	30. วันที่ติดต่อคือวัน mon ใช่หรือไม่ (x_{91}) (ใช่ = 1, ไม่ใช่ = 0) 31. วันที่ติดต่อคือวัน thu ใช่หรือไม่ (x_{92}) (ใช่ = 1, ไม่ใช่ = 0) 32. วันที่ติดต่อคือวัน wed ใช่หรือไม่ (x_{93}) (ใช่ = 1, ไม่ใช่ = 0) 33. วันที่ติดต่อคือวัน tue ใช่หรือไม่ (x_{94}) (ใช่ = 1, ไม่ใช่ = 0) 34. วันที่ติดต่อคือวัน fri ใช่หรือไม่ (x_{95}) (ใช่ = 1, ไม่ใช่ = 0)
10. Poutcome (x_{10}) (failure, success, nonexistent)	35. แปลงให้อยู่ในรูปปรกติโดยมีค่า 0 ถึง 1 (x_{10}) (max = 3, min = 1)
11. Duration (x_{11})	36. แปลงให้อยู่ในรูปปรกติโดยมีค่า 0 ถึง 1 (x_{11}) (max = 4,918, min = 0)
12. Age (x_{12})	37. แปลงให้อยู่ในรูปปรกติโดยมีค่า 0 ถึง 1 (x_{12}) (max = 95, min = 17)
13. Campaign (x_{13})	38. แปลงให้อยู่ในรูปปรกติโดยมีค่า 0 ถึง 1 (x_{13}) (max = 43, min = 1)
14. Pdays (x_{14})	39. แปลงให้อยู่ในรูปปรกติโดยมีค่า 0 ถึง 1 (x_{14})

ตัวแปรอิสระก่อนการแปลง	ตัวแปรอิสระหลังการแปลง
	(max = 999, min = 0)
15. Previous (x_{15})	40. แปลงให้อยู่ในรูปปรกติโดยมีค่า 0 ถึง 1 (x_{15}) (max = 7, min = 0)
16. Emp.var.rate (x_{16})	41. แปลงให้อยู่ในรูปปรกติโดยมีค่า 0 ถึง 1 (x_{16}) (max = 1.4, min = -3.4)
17. Cons.price.idx (x_{17})	42. แปลงให้อยู่ในรูปปรกติโดยมีค่า 0 ถึง 1 (x_{17}) (max = 94.767, min = 92.201)
18. Cons.conf.idx (x_{18})	43. แปลงให้อยู่ในรูปปรกติโดยมีค่า 0 ถึง 1 (x_{18}) (max = -26.9, min = -50.8)
19. Euribor3m (x_{19})	44. แปลงให้อยู่ในรูปปรกติโดยมีค่า 0 ถึง 1 (x_{19}) (max = 5.045, min = 0.634)
20. Nr.employed (x_{20})	45. แปลงให้อยู่ในรูปปรกติโดยมีค่า 0 ถึง 1 (x_{20}) (max = 5228.1, min = 4963.6)

3.3.2 ชุดข้อมูลสายพันธุ์ข้าว

3.3.2.1 ศึกษารายละเอียดชุดข้อมูล

ชุดข้อมูลสายพันธุ์ คือชุดข้อมูลที่เก็บรวบรวมข้อมูลของข้าว 2 สายพันธุ์ได้แก่ข้าวสายพันธุ์ Osmancik (ในที่นี้จะแทนด้วย No) และ Cammeo (ในที่นี้จะแทนด้วย Yes) ที่ปลูกในประเทศตุรกี โดยมีรายละเอียดตัวแปรดังตารางที่ 8

ตารางที่ 8 แสดงรายละเอียดชุดข้อมูลสายพันธุ์ข้าว

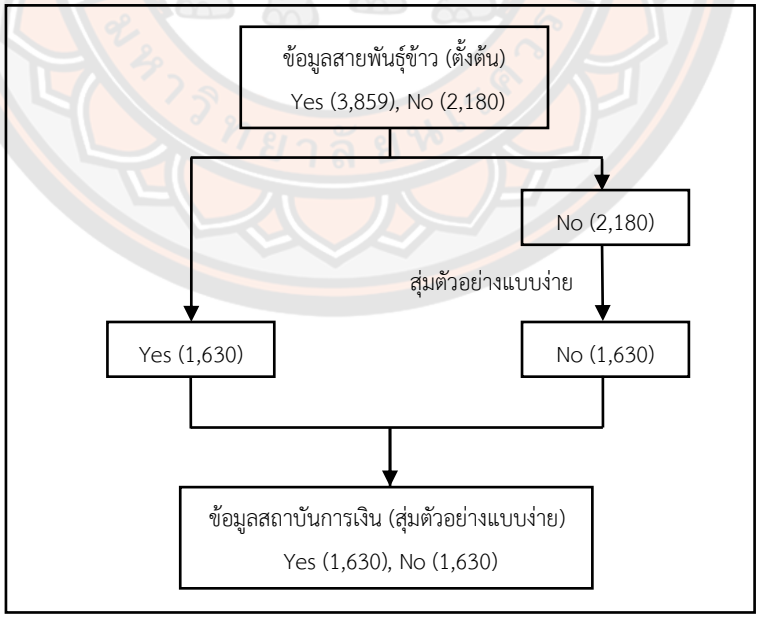
ตัวแปร	รายละเอียด
Area	จำนวนพิกเซลภายในขอบเขตของเมล็ดข้าว (ตัวแปรเชิงปริมาณ)
Perimeter	เส้นรอบวงโดยการคำนวณระยะห่างระหว่างพิกเซลรอบขอบเขตของเมล็ดข้าว (ตัวแปรเชิงปริมาณ)
Major Axis Length	เส้นที่ยาวที่สุดที่สามารถวาดบนเมล็ดข้าวได้คือระยะแกนหลักให้ (ตัวแปรเชิงปริมาณ)
Minor Axis Length	เส้นที่สั้นที่สุดที่สามารถวาดบนเมล็ดข้าวได้ คือ ระยะแกนเล็ก (ตัวแปรเชิงปริมาณ)

ตัวแปร	รายละเอียด
Eccentricity	วัดความกลมของวงรีซึ่งมีโมเมนต์เดียวกับเมล็ดข้าว (ตัวแปรเชิงปริมาณ)
Convex Area	จำนวนพิกเซลของเปลือกนูนที่เล็กที่สุดของบริเวณที่เกิดจากเมล็ดข้าว (ตัวแปรเชิงปริมาณ)
Extent	อัตราส่วนของพื้นที่ที่เกิดจากเมล็ดข้าวเป็นพิกเซลของกรอบ (ตัวแปรเชิงปริมาณ)
Y	ข้าวสายพันธุ์ Cammeo และ Osmancik (ตัวแปรเชิงคุณภาพ)

3.3.2.2 การปรับปรุงชุดข้อมูลอสมดุลให้สมดุล

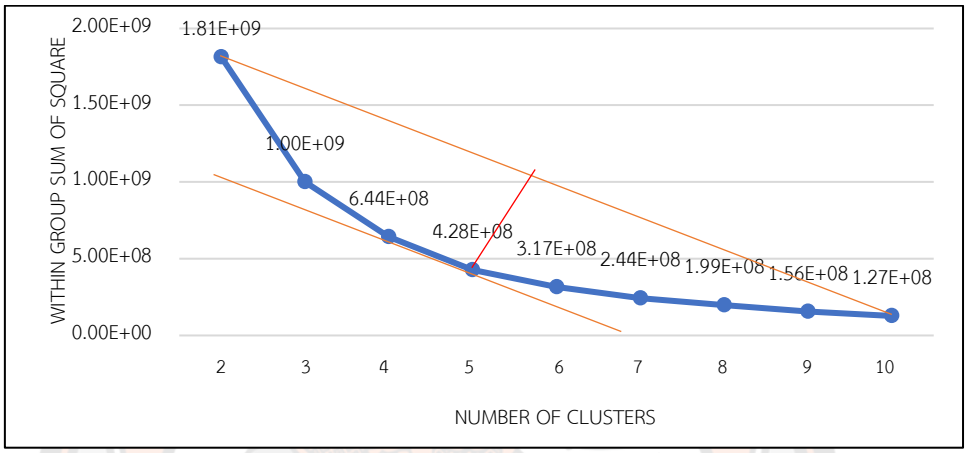
ในงานวิจัยนี้ผู้วิจัยจะใช้วิธีสุ่มลด โดยใช้การสุ่มตัวอย่างแบบง่ายและการแบ่งกลุ่มข้อมูลแบบเคมีน โดยมีรายละเอียดดังต่อไปนี้

1. การสุ่มตัวอย่างแบบง่าย คือการสุ่มตัวอย่างโดยกำหนดให้ทุกหน่วยมีโอกาสถูกเลือกเท่า ๆ กันแบบไม่แทนที่ โดยในที่นี้ทำการสุ่มลดจำนวนข้อมูลของกลุ่ม No ให้มีขนาดเท่ากับกลุ่ม Yes แสดงดังในภาพที่ 20

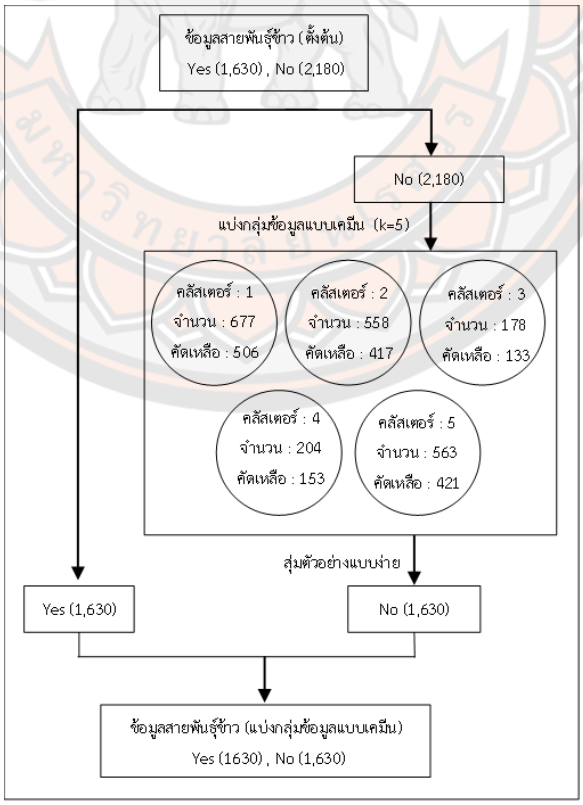


ภาพที่ 20 การปรับปรุงชุดข้อมูลอสมดุลให้สมดุลด้วยวิธีการสุ่มตัวอย่างแบบง่ายของชุดข้อมูลสายพันธุ์ข้าว

2. การแบ่งกลุ่มข้อมูลแบบเคมีน โดยจะพิจารณาลดจำนวนข้อมูลของกลุ่ม No ภายใต้ชุดสายพันธุ์ข้าวพบว่าค่า k ที่เหมาะสมมีค่าเท่ากับ 5 ดังภาพที่ 21 จากนั้นทำการสุ่มตัวอย่างแบบง่ายตามสัดส่วนของสมาชิกแต่ละคลัสเตอร์ ยกตัวอย่างเช่น ทำการสุ่มตัวอย่างแบบง่ายจากคลัสเตอร์ที่ 1 จำนวน $\left(\frac{677}{2,180}\right)\times 1630 = 506$ และทำเช่นเดียวกันกับคลัสเตอร์ที่เหลือดังภาพที่ 22



ภาพที่ 21 ภายใต้ชุดข้อมูลสายพันธุ์ข้าวค่า k ที่เหมาะสมคือ 5 โดยวิธีของ Elbow Method



ภาพที่ 22 การปรับปรุงชุดข้อมูลอสมดุลให้สมดุลด้วยเทคนิคการแบ่งกลุ่มข้อมูลแบบเคมีนของชุดข้อมูลสายพันธุ์ข้าว เมื่อค่า k ที่เหมาะสมเท่ากับ 5

3.3.2.3 ทำการแปลงตัวแปรอิสระให้เป็นปรกติ

ในการพัฒนาตัวแบบการจำแนกในงานวิจัยนี้ผู้วิจัยจะมีการแปลงตัวแปรอิสระ โดยตัวแปรเชิงปริมาณจะปรับให้เป็นปรกติ มีค่าระหว่าง 0 – 1 โดยมีรายละเอียดดังตารางต่อไปนี้

ตัวแปรอิสระก่อนการแปลง	ตัวแปรอิสระหลังการแปลง
1. Area (x_1)	1. แปลงให้อยู่ในรูปปรกติโดยมีค่า 0 ถึง 1 (x_1) (max = 18,913, min = 7,551)
2. Perimeter (x_2)	2. แปลงให้อยู่ในรูปปรกติโดยมีค่า 0 ถึง 1 (x_2) (max = 548.446, min = 359.1)
3. Major Axis Length (x_3)	3. แปลงให้อยู่ในรูปปรกติโดยมีค่า 0 ถึง 1 (x_3) (max = 239.0105, min = 145.2645)
4. Minor Axis Length (x_4)	4. แปลงให้อยู่ในรูปปรกติโดยมีค่า 0 ถึง 1 (x_4) (max = 107.5425, min = 59.5324)
5. Eccentricity (x_5)	5. แปลงให้อยู่ในรูปปรกติโดยมีค่า 0 ถึง 1 (x_5) (max = 0.948, min = 0.7772)
6. Convex Area (x_6)	6. แปลงให้อยู่ในรูปปรกติโดยมีค่า 0 ถึง 1 (x_6) (max = 19,099, min = 7,723)
7. Extent (x_7)	7. แปลงให้อยู่ในรูปปรกติโดยมีค่า 0 ถึง 1 (x_7) (max = 0.86105, min = 0.497413)

3.3.3 ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล

3.3.3.1 ศึกษารายละเอียดชุดข้อมูล

ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล คือชุดข้อมูลที่ถูกเก็บรวบรวมโดยบริษัทแห่งหนึ่งซึ่งเป็นบริษัทที่ทำอยู่ในสายงานข้อมูลขนาดใหญ่และวิทยาศาสตร์ข้อมูล โดยจัดอบรมหลักสูตรมีเป้าหมายเพื่อต้องการทราบว่าหลังการอบรมมีผู้อบรมท่านใดที่สนใจจะสมัครหรือเปลี่ยนงานมาเป็นนักวิทยาศาสตร์ข้อมูล โดยมีรายละเอียดตัวแปรดังตารางที่ 9

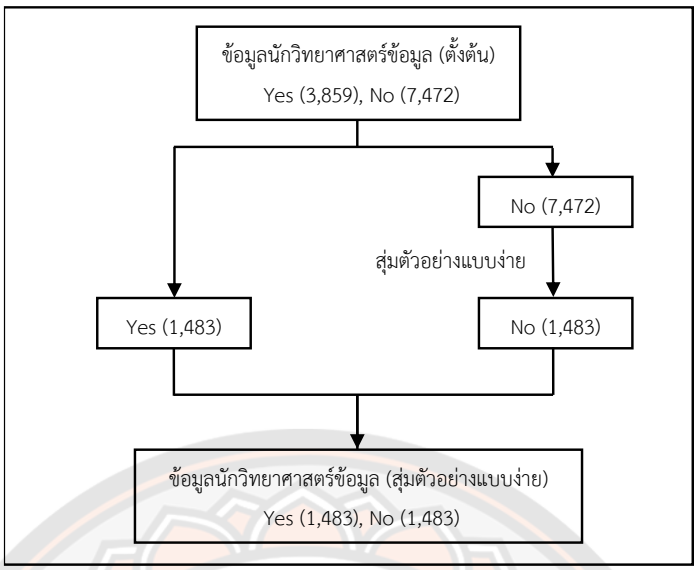
ตารางที่ 9 แสดงรายละเอียดชุดข้อมูลนักวิทยาศาสตร์ข้อมูล

ตัวแปร	รายละเอียด
Gender	เพศ (ตัวแปรเชิงคุณภาพ มาตรนามบัญญัติ)
Relevent_experience	ประสบการณ์ที่เกี่ยวข้องของผู้สมัคร (ตัวแปรเชิงคุณภาพ มาตรเรียงลำดับ)
Enrolled_university	ประเภทของหลักสูตรของมหาวิทยาลัยที่ลงทะเบียน (ตัวแปรเชิงคุณภาพ มาตรนามบัญญัติ)
Education_level	ระดับการศึกษา (ตัวแปรเชิงคุณภาพ มาตรเรียงลำดับ)
Major_discipline	วิชาเอกของผู้สมัคร (ตัวแปรเชิงคุณภาพ มาตรนามบัญญัติ)
Experience	ประสบการณ์ทั้งหมดของผู้สมัคร (ตัวแปรเชิงคุณภาพ มาตรเรียงลำดับ)
Company_size	จำนวนพนักงานในบริษัทของนายจ้างปัจจุบัน (ตัวแปรเชิงคุณภาพ มาตรเรียงลำดับ)
Company_type	ประเภทบริษัทปัจจุบัน (ตัวแปรเชิงคุณภาพ มาตรนามบัญญัติ)
Last_new_job	ความแตกต่างระหว่างงานก่อนหน้ากับงานปัจจุบัน (ตัวแปรเชิงคุณภาพ มาตรนามบัญญัติ)
Training_hours	ชั่วโมงการฝึกอบรม (ตัวแปรเชิงปริมาณ)
Y	ต้องการเป็นนักวิทยศาสตร์ข้อมูลหรือไม่ (ตัวแปรเชิงคุณภาพ)

3.3.3.2 การปรับปรุงชุดข้อมูลอสมดุลให้สมดุล

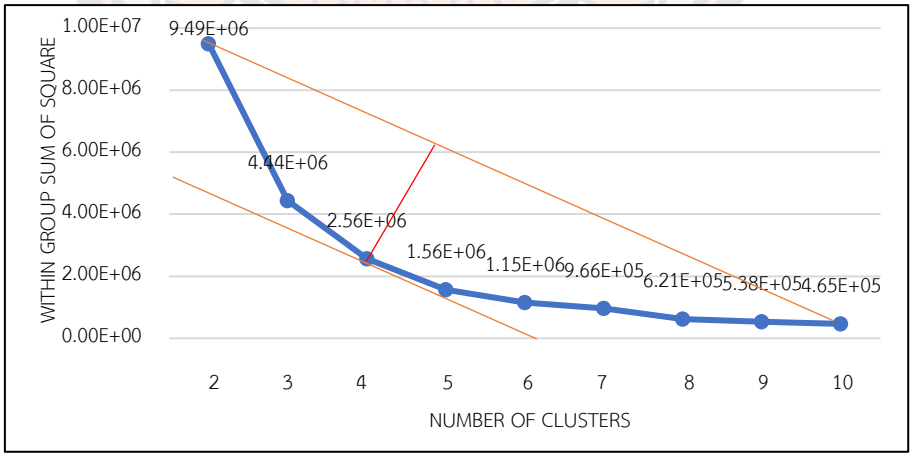
ในงานวิจัยนี้ผู้วิจัยจะใช้วิธีสุ่มลด โดยใช้เทคนิคการสุ่มตัวอย่างแบบง่ายและเทคนิคการแบ่งกลุ่มข้อมูลแบบเคมีน โดยมีรายละเอียดดังต่อไปนี้

1. การสุ่มตัวอย่างแบบง่าย คือการสุ่มตัวอย่างโดยกำหนดให้ทุกหน่วยมีโอกาสถูกเลือกเท่า ๆ กันแบบไม่แทนที่ โดยในที่นี้ทำการสุ่มลดจำนวนกลุ่ม No ให้มีขนาดเท่ากับกลุ่ม Yes แสดงดังในภาพที่ 23

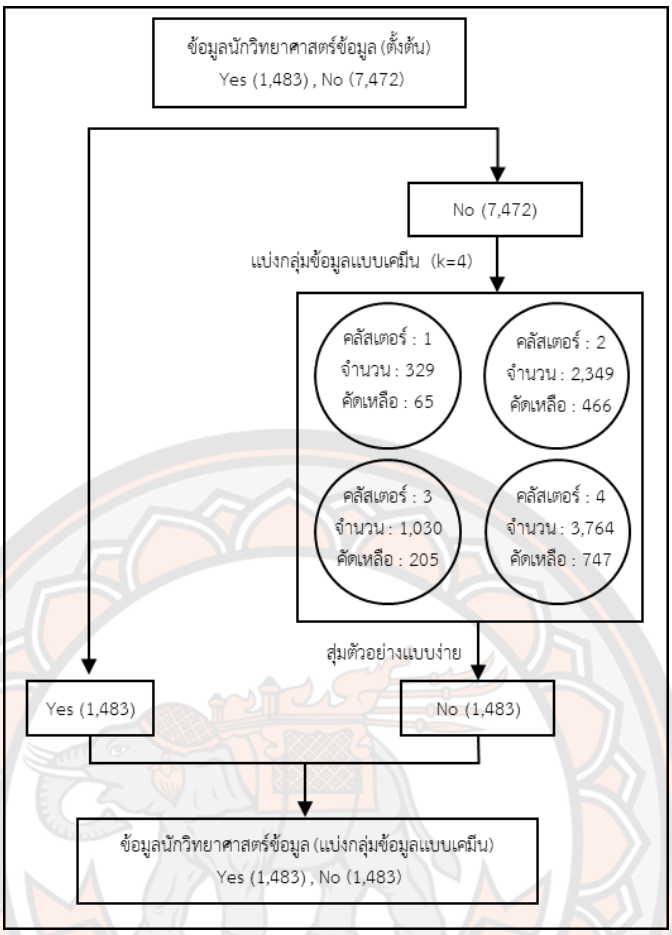


ภาพที่ 23 โครงสร้างการปรับปรุงชุดข้อมูลอสมดุลให้สมดุลด้วยวิธีการสุ่มตัวอย่างแบบง่าย ภายใต้ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล

2. เทคนิคการแบ่งกลุ่มข้อมูลแบบเคมีน โดยจะพิจารณาลดจำนวนของกลุ่ม No ภายใต้ชุดข้อมูลสถาบันการเงิน พบว่าค่า k ที่เหมาะสมมีค่าเท่ากับ 4 ดังภาพที่ 24 จากนั้นทำการสุ่มตัวอย่างแบบง่ายตามสัดส่วนของสมาชิกแต่ละคลัสเตอร์ ยกตัวอย่างเช่น ทำการสุ่มตัวอย่างแบบง่ายจากคลัสเตอร์ที่ 1 จำนวน $\left(\frac{329}{7,472}\right)\times 1,483 = 65$ และทำเช่นเดียวกันกับคลัสเตอร์ที่เหลือดังภาพที่ 25



ภาพที่ 24 ภายใต้ชุดข้อมูลนักวิทยาศาสตร์ข้อมูลค่า k ที่เหมาะสมคือ 4 โดยวิธีของ Elbow Method



ภาพที่ 25 โครงสร้างปรับปรุงชุดข้อมูลอสมดุลให้สมดุลด้วยเทคนิคการแบ่งกลุ่มข้อมูลแบบเคมีนภายใต้ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล เมื่อค่า k ที่เหมาะสมเท่ากับ 4

3.3.3.3 ทำการแปลงตัวแปรอิสระให้เป็นปรกติและเป็นตัวแปรดัมมี่

ในการพัฒนาตัวแบบการจำแนกในงานวิจัยนี้ผู้วิจัยจะมีการแปลงตัวแปรอิสระ โดยตัวแปรเชิงปริมาณและตัวแปรเชิงคุณภาพมาตรเรียงลำดับจะปรับให้เป็นปรกติ มีค่าระหว่าง 0 – 1 และตัวแปรเชิงคุณภาพมาตรนามบัญญัติจะแปลงให้เป็นตัวแปรดัมมี่แบบไบนารี่ โดยมีรายละเอียดดังตารางต่อไปนี้

ตัวแปรอิสระก่อนการแปลง	ตัวแปรอิสระหลังการแปลง
1. Gender (x_1) (Male, Female, Other)	1. เพศคือ Male ใช่หรือไม่ (x_{11}) (ใช่ =1, ไม่ใช่ = 0) 2. เพศคือ Female ใช่หรือไม่ (x_{12}) (ใช่ =1, ไม่ใช่ = 0) 3. เพศคือ Other ใช่หรือไม่ (x_{13}) (ใช่ =1, ไม่ใช่ = 0)
2. Relevent_experience (x_2) (Has relevent experience, No relevent experience)	4. มีประสบการณ์ที่เกี่ยวข้องใช่หรือไม่ (x_2) (ใช่ =1, ไม่ใช่ = 0)
3. Enrolled_university (x_3) (no_enrollment, Part time course, Full time course)	5. หลักสูตรคือ no_enrollment ใช่หรือไม่ (x_{31}) (ใช่ =1, ไม่ใช่ = 0) 6. หลักสูตรคือ Part time course ใช่หรือไม่ (x_{32}) (ใช่ =1, ไม่ใช่ = 0) 7. หลักสูตรคือ Full time course ใช่หรือไม่ (x_{33}) (ใช่ =1, ไม่ใช่ = 0)
4. Education_level (x_4) (Phd, , Masters, Graduate)	8. แปลงให้อยู่ในรูปปรกติโดยมีค่า 0 ถึง 1 (x_4) (max = 3, min = 1)
5. Major_discipline (x_5) (Arts, Business Degree, Humanities, No Major, STEM, Other)	9. วิชาเอกคือ Arts ใช่หรือไม่ (x_{51}) (ใช่ =1, ไม่ใช่ = 0) 10. วิชาเอกคือ Business Degree ใช่หรือไม่ (x_{52}) (ใช่ =1, ไม่ใช่ = 0) 11. วิชาเอกคือ Humanities ใช่หรือไม่ (x_{53}) (ใช่ =1, ไม่ใช่ = 0) 12. วิชาเอกคือ No Major ใช่หรือไม่ (x_{54}) (ใช่ =1, ไม่ใช่ = 0) 13. วิชาเอกคือ STEM ใช่หรือไม่ (x_{55}) (ใช่ =1, ไม่ใช่ = 0) 14. วิชาเอกคือ Other ใช่หรือไม่ (x_{56}) (ใช่ =1, ไม่ใช่ = 0)

ตัวแปรอิสระก่อนการแปลง	ตัวแปรอิสระหลังการแปลง
6. Experience (x_6) (0, 1to5, 6to10, 11to15, 16to20, 20up)	15. แปลงให้อยู่ในรูปปรกติโดยมีค่า 0 ถึง 1 (x_6) (max = 6, min = 1)
7. Company_size (x_7) (<10, 10to49, 50to99, 100to500, 500to999, 1000to4999, 5000to9999, 10000up)	16. แปลงให้อยู่ในรูปปรกติโดยมีค่า 0 ถึง 1 (x_7) (max = 8, min = 1)
8. Company_type (x_8) (Early-Stage Startup, Funded Startup, NGO, Public Sector, Pvt Ltd, Other)	17. ประเภทของบริษัทคือ Early-Stage Startup ใช่หรือไม่ (x_{81}) (ใช่ =1, ไม่ใช่ = 0) 18. ประเภทของบริษัทคือ Funded Startup ใช่หรือไม่ (x_{82}) (ใช่ =1, ไม่ใช่ = 0) 19. ประเภทของบริษัทคือ NGO ใช่หรือไม่ (x_{83}) (ใช่ =1, ไม่ใช่ = 0) 20. ประเภทของบริษัทคือ Public Sector ใช่หรือไม่ (x_{84}) (ใช่ =1, ไม่ใช่ = 0) 21. ประเภทของบริษัทคือ Pvt Ltd ใช่หรือไม่ (x_{85}) (ใช่ =1, ไม่ใช่ = 0) 22. ประเภทของบริษัทคือ Other ใช่หรือไม่ (x_{86}) (ใช่ =1, ไม่ใช่ = 0)
9. Last_new_job (x_9) (0, 1, 2, 3, 4, 4up)	23. ความแตกต่างระหว่างงานก่อนหน้าและงานปัจจุบันเท่ากับ 0 ปีใช่หรือไม่ (x_{91}) (ใช่ =1, ไม่ใช่ = 0) 24. ความแตกต่างระหว่างงานก่อนหน้าและงานปัจจุบันเท่ากับ 1 ปีใช่หรือไม่ (x_{92}) (ใช่ =1, ไม่ใช่ = 0) 25. ความแตกต่างระหว่างงานก่อนหน้าและงานปัจจุบันเท่ากับ 2 ปีใช่หรือไม่ (x_{93})

ตัวแปรอิสระก่อนการแปลง	ตัวแปรอิสระหลังการแปลง
	(ใช่ =1, ไม่ใช่ = 0) 26. ความแตกต่างระหว่างงานก่อนหน้าและงานปัจจุบันเท่ากับ 3 ปีใช่หรือไม่ (x_{94}) (ใช่ =1, ไม่ใช่ = 0) 27. ความแตกต่างระหว่างงานก่อนหน้าและงานปัจจุบันเท่ากับ 4 ปีใช่หรือไม่ (x_{95}) (ใช่ =1, ไม่ใช่ = 0) 28. ความแตกต่างระหว่างงานก่อนหน้าและงานปัจจุบันมากกว่า 4 ปีใช่หรือไม่ (x_{96}) (ใช่ =1, ไม่ใช่ = 0)
10. Training_hours (x_{10})	29. แปลงให้อยู่ในรูปปรกติโดยมีค่า 0 ถึง 1 (x_{10}) (max = 336, min = 1)

หลังจากการศึกษารายละเอียดชุดข้อมูล การปรับปรุงชุดข้อมูลอสมดุลให้สมดุลและการแปลงตัวแปรอิสระให้เป็นปรกติและเป็นตัวแปรดัมมี่ ดังนั้นจะมีชุดข้อมูลทั้งหมด 9 ชุด ซึ่งแต่ละชุดมีรายละเอียดดังตารางที่ 10

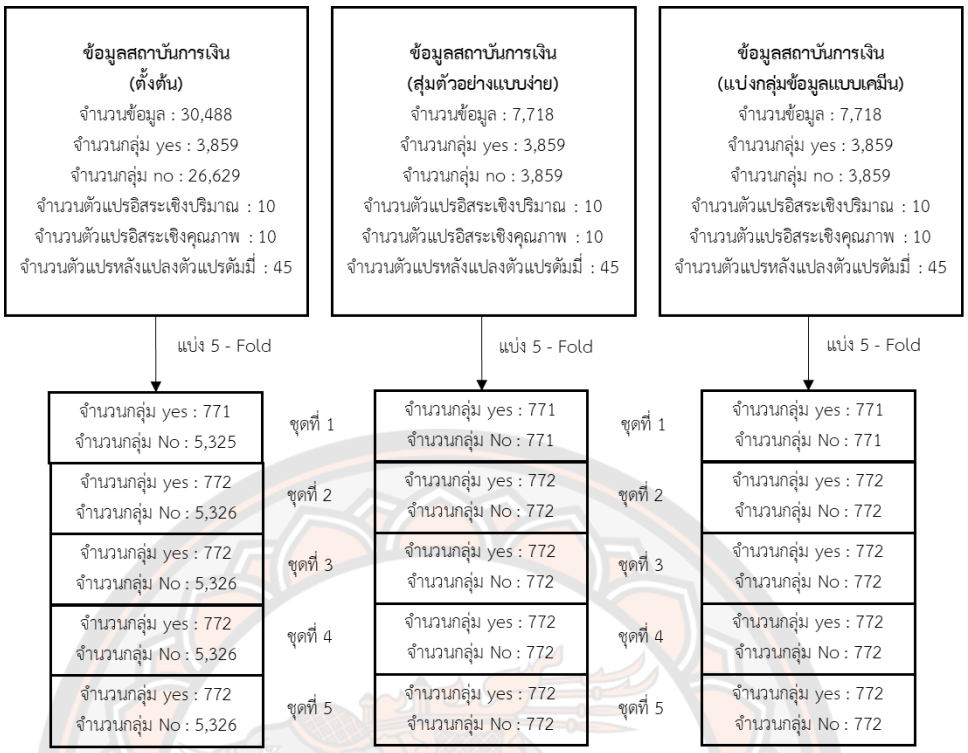
ตารางที่ 10 รายละเอียดชุดข้อมูลทั้งหมดที่ใช้ในงานวิจัย

ชื่อชุดข้อมูล	จำนวนตัวแปรอิสระ		ค่าตัวแปรตาม		จำนวนข้อมูล	
	ตัวแปรดัมมี่	ตัวแปรปรกติ	No	Yes	จำนวน	สัดส่วนความไม่สมดุล
1. สถาบันการเงิน (ตั้งต้น)	35	10	26,629	3,859	30,488	1:6.7
2. สายพันธุ์ข้าว (ตั้งต้น)	0	7	2,180	1,630	3,810	1:1.3
3. นักวิทยาศาสตร์ข้อมูล (ตั้งต้น)	28	1	7,472	1,483	8,599	1:4.9
4. สถาบันการเงิน (สุ่มตัวอย่างแบบง่าย)	35	10	3,859	3,859	7,718	1:1

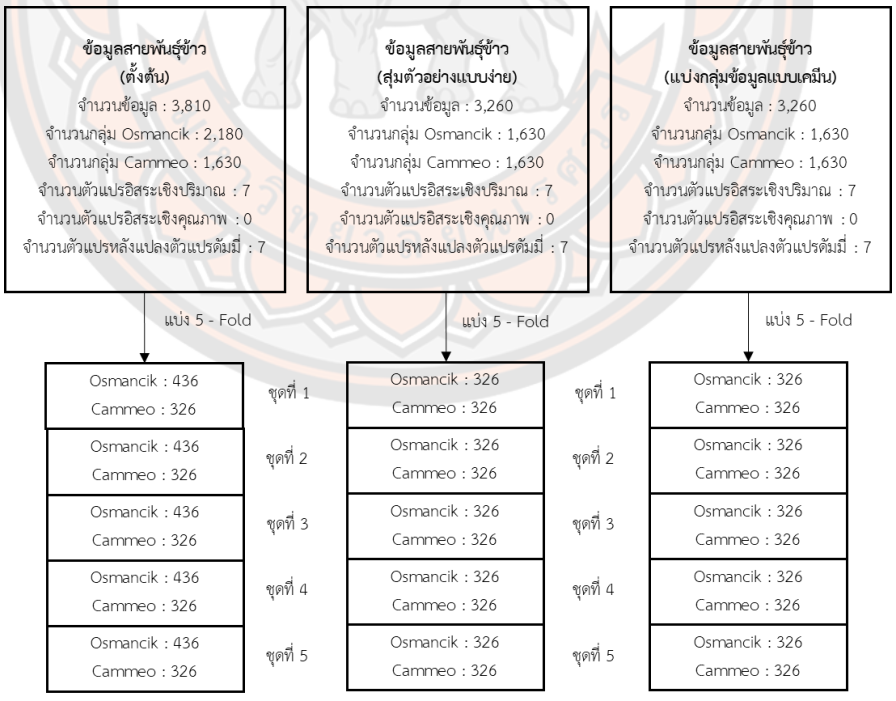
ชื่อชุดข้อมูล	จำนวนตัวแปรอิสระ		ค่าตัวแปรตาม		จำนวนข้อมูล	
	ตัวแปรดัมมี่	ตัวแปรปรกติ	No	Yes	จำนวน	สัดส่วนความไม่สมดุล
5. สายพันธุ์ข้าว (สุ่มตัวอย่างแบบง่าย)	0	7	1,630	1,630	3,260	1:1
6. นักวิทยาศาสตร์ข้อมูล (การสุ่มตัวอย่างแบบง่าย)	28	1	1,483	1,483	2,966	1:1
7. สถาบันการเงิน (แบ่งกลุ่มข้อมูลแบบเคมีน)	35	10	3,859	3,859	7,718	1:1
8. สายพันธุ์ข้าว (แบ่งกลุ่มข้อมูลแบบเคมีน)	0	7	1,630	1,630	3,260	1:1
9. นักวิทยาศาสตร์ข้อมูล (แบ่งกลุ่มข้อมูลแบบเคมีน)	28	1	1,483	1,483	2,966	1:1

3.3.4 การแบ่งชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ

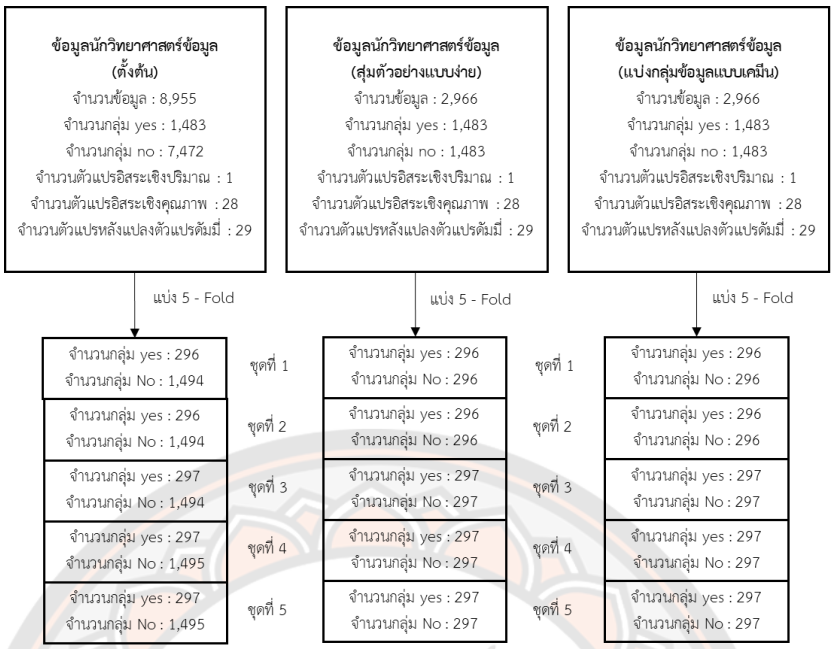
ในหัวข้อนี้ผู้วิจัยจะนำชุดข้อมูล 9 ชุด (จากตารางที่ 10) นำมาแบ่งเป็นชุดมูลเรียนรู้และชุดข้อมูลทดสอบด้วยหลักการ 5-Fold เพื่อเตรียมในการสร้างตัวแบบการจำแนก โดยแต่ละชุดข้อมูลมีรายละเอียดดังภาพที่ 26 - 28



ภาพที่ 26 การแบ่งชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ ภายใต้ชุดข้อมูลสถาบันการเงิน



ภาพที่ 27 การแบ่งชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ ภายใต้ชุดข้อมูลสายพันธุ์ข้าว



ภาพที่ 28 การแบ่งชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ ภายใต้ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล

3.4 ขั้นตอนพัฒนาตัวแบบการจำแนก

ในหัวข้อนี้ผู้วิจัยจะนำชุดข้อมูลจำนวน 9 ชุดซึ่งประกอบด้วย ชุดข้อมูลสถาบันการเงิน (ตั้งต้น) ชุดข้อมูลสถาบันการเงิน (สุ่มตัวอย่างแบบง่าย) ชุดข้อมูลสถาบันการเงิน (แบ่งกลุ่มข้อมูลแบบเคมีน) ชุดข้อมูลสายพันธุ์ข้าว (ตั้งต้น) ชุดข้อมูลสายพันธุ์ข้าว (สุ่มตัวอย่างแบบง่าย) ชุดข้อมูลสายพันธุ์ (แบ่งกลุ่มข้อมูลแบบเคมีน) ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (ตั้งต้น) ชุดข้อมูลนักวิทยาศาสตร์ (สุ่มตัวอย่างแบบง่าย) และชุดข้อมูลนักวิทยาศาสตร์ (แบ่งกลุ่มข้อมูลแบบเคมีน) มาสร้างตัวแบบการจำแนก

พารามิเตอร์ของการสร้างตัวแบบการจำแนกด้วยการวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์ เทคนิคนาอีฟเบย์ ต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 และเทคนิคป่าสุ่มจะใช้การกำหนดแบบพื้นฐานยกเว้นเทคนิคโครงข่ายประสาทเทียมที่ต้องมีการกำหนดพารามิเตอร์แบบต่าง ๆ เพื่อให้ได้ตัวแบบที่มีประสิทธิภาพ

การกำหนดพารามิเตอร์เทคนิคโครงข่ายประสาทเทียม

เทคนิคโครงข่ายประสาทเทียมจำเป็นต้องมีการกำหนดพารามิเตอร์ที่สำคัญได้แก่ อัตราการเรียนรู้ ค่าโมเมนตัม จำนวนชั้นซ่อนและจำนวนโหนดในชั้นซ่อน โดยในงานวิจัยนี้จะกำหนดพารามิเตอร์ดังต่อไปนี้

1. อัตราการเรียนรู้ กำหนดเท่ากับ 0.1, 0.2
2. ค่าโมเมนตัม กำหนดเท่ากับ 0.1, 0.2

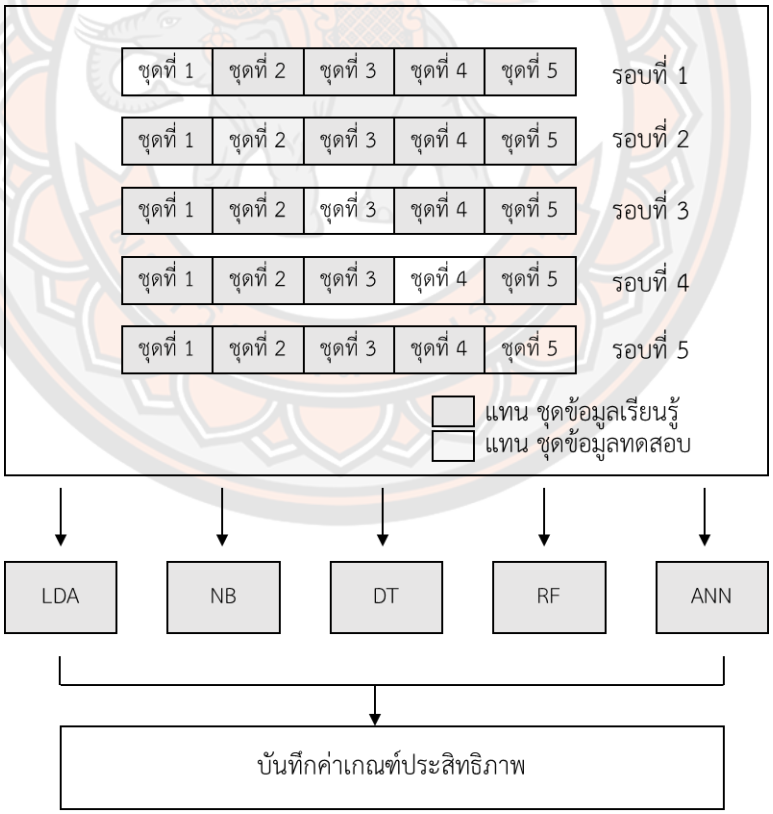
3. จำนวนชั้นซ่อนเท่ากับ 1 และกำหนดจำนวนโหนดในชั้นซ่อนโดยพิจารณาจากโหนดในข้อมูลเข้าและโหนดในชั้นผลลัพธ์ โดยคำนวณสูตรต่อไปนี้

$$\text{จำนวนโหนดในชั้นซ่อน} = \frac{\text{จำนวนโหนดในชั้นนำเข้า} + \text{จำนวนโหนดในชั้นผลลัพธ์}}{2}$$

ในงานวิจัยนี้กำหนดจำนวนโหนดในขั้นผลลัพธ์ทุกชุดข้อมูลเท่ากับ 2 เมื่อได้จำนวนโหนดในขั้นซ่อนแล้วจะกำหนดจำนวนโหนดในชั้นซ่อนเพิ่มเติมอีก 4 ค่า (± 2) ดังนั้นจะกำหนดจำนวนโหนดในชั้นซ่อนแต่ละชุดข้อมูลดังต่อไปนี้

1. ชุดข้อมูลสถาบันการเงิน กำหนดเท่ากับ 21, 22, **23**, 24, 25
2. ชุดข้อมูลสายพันธุ์ข้าว กำหนดเท่ากับ 2, 3, **4**, 5, 6
3. ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล กำหนดเท่ากับ 13, 14, **15**, 16, 17

ในขั้นตอนต่อไปคือการนำชุดข้อมูลทั้ง 9 ชุดทำมาสร้างตัวแบบการจำแนกและบันทึกค่าเกณฑ์ประสิทธิภาพ ดังภาพที่ 29



ภาพที่ 29 แสดงการพัฒนาตัวแบบการจำแนก

3.5 การประเมินประสิทธิภาพ

โดยเกณฑ์ที่ใช้พิจารณาประกอบด้วย

1. ค่าความแม่น **(Accuracy)** คำนวณดังสมการที่ 1
2. ค่าความเที่ยง **(Precision)**
   2.1 ค่าความเที่ยงของกลุ่ม Yes คำนวณดังสมการที่ 2
   2.2 ค่าความเที่ยงของกลุ่ม No คำนวณดังสมการที่ 3
3. ค่าการเรียกคืน **(Recall)**
   3.1 ค่าการเรียกคืนของกลุ่ม Yes คำนวณดังสมการที่ 4
   3.2 ค่าการเรียกคืนของกลุ่ม No คำนวณดังสมการที่ 5
4. ค่าประสิทธิภาพ **(F-measure)**
   4.1 ค่าประสิทธิภาพของกลุ่ม Yes คำนวณดังสมการที่ 6
   4.2 ค่าประสิทธิภาพของกลุ่ม No คำนวณดังสมการที่ 7

บทที่ 4
ผลการวิจัย

ในบทนี้ผู้วิจัยจะนำเสนอผลการวิจัยที่ได้จากการสร้างและพัฒนาตัวแบบการจำแนก ซึ่งจะแสดงผลลัพธ์การสร้างตัวแบบการจำแนกประกอบด้วย การวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์ เทคนิคนาอีฟเบย์ ต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 เทคนิคป่าสุ่ม โครงข่ายประสาทเทียม ซึ่งทำงานภายใต้ชุดข้อมูลจำนวน 9 ชุดประกอบด้วย ชุดข้อมูลสถาบันการเงิน (ตั้งต้น) ชุดข้อมูลสายพันธุ์ข้าว (ตั้งต้น) ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (ตั้งต้น) ชุดข้อมูลสถาบันการเงินที่ (สุ่มตัวอย่างแบบง่าย) ชุดข้อมูลสายพันธุ์ข้าว (สุ่มตัวอย่างแบบง่าย) ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (สุ่มตัวอย่างแบบง่าย) ชุดข้อมูลสถาบันการเงิน (แบ่งกลุ่มข้อมูลแบบเคมีน) ชุดข้อมูลสายพันธุ์ข้าว (แบ่งกลุ่มข้อมูลแบบเคมีน) และชุดข้อมูลนักวิทยาศาสตร์ (แบ่งกลุ่มข้อมูลแบบเคมีน) โดยมีรายละเอียดดังต่อไปนี้

1. ผลการเปรียบเทียบประสิทธิภาพการจำแนกภายใต้ชุดข้อมูลตั้งต้น
2. ผลการเปรียบเทียบประสิทธิภาพการจำแนกภายใต้ชุดข้อมูลสุ่มตัวอย่างแบบง่าย
3. ผลการเปรียบเทียบประสิทธิภาพการจำแนกภายใต้ชุดข้อมูลแบ่งกลุ่มข้อมูลแบบเคมีน
4. ผลการเปรียบเทียบประสิทธิภาพการจำแนกภายใต้ชุดข้อมูลสุ่มตัวอย่างแบบง่ายกับชุดข้อมูลแบ่งกลุ่มข้อมูลแบบเคมีน

4.1 ผลการเปรียบเทียบประสิทธิภาพการจำแนกภายใต้ชุดข้อมูลตั้งต้น

ในหัวข้อนี้ผู้วิจัยจะนำเสนอผลการเปรียบเทียบประสิทธิภาพตัวแบบการจำแนกภายใต้ชุดข้อมูลสถาบันการเงิน (ตั้งต้น) ชุดข้อมูลสายพันธุ์ข้าว (ตั้งต้น) และชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (ตั้งต้น) ดังตารางที่ 11 – 13 ตามลำดับ

จากตารางที่ 11 - 22 แสดงผลลัพธ์การจำแนกภายใต้ชุดข้อมูลต่าง ๆ โดยมีรายละเอียดดังต่อไปนี้ LDA แทนตัวแบบการจำแนกจากเทคนิคการวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์ NB แทนตัวแบบการจำแนกจากเทคนิคนาอีฟเบย์ DT แทนตัวแบบการจำแนกจากเทคนิคต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 RF แทนตัวแบบการจำแนกจากเทคนิคป่าสุ่ม ANN แทนตัวแบบการจำแนกจากเทคนิคโครงข่ายประสาทเทียม Pre แทนค่าความเที่ยง Re แทนค่าการเรียกคืน F แทนค่าประสิทธิภาพ Acc แทนค่าความแม่นและ Rank แทนการเรียงลำดับตัวแบบที่มีประสิทธิภาพมากที่สุดเช่นค่า Rank เท่ากับ 1 หมายถึงตัวแบบการจำแนกที่มีประสิทธิภาพมากที่สุดและ Rank เท่ากับ 5 หมายถึงตัวแบบการจำแนกที่มีประสิทธิภาพต่ำที่สุด

ตารางที่ 11 แสดงผลลัพธ์การจำแนกภายใต้ชุดข้อมูลสถาบันการเงิน (ตั้งต้น)

	Yes			No			Acc
	Pre	Re	F	Pre	Re	F	
LDA	0.625	0.503	0.557	0.930	0.956	0.943	0.899
NB	0.410	**0.588**	0.483	0.936	0.877	0.906	0.840
DT	0.589	0.522	0.554	0.932	0.947	0.939	0.893
RF	**0.665**	0.480	**0.557**	0.927	**0.965**	**0.946**	0.903
ANN (0.2,0.1,21)	0.654	0.520	**0.577**	**0.933**	0.960	**0.946**	**0.904**

ตัวเลขในวงเล็บของตัวแบบ ANN แทน ค่าโมเมนตัม อัตราการเรียนรู้และจำนวนโหนดในชั้นซ่อน ตามลำดับ

จากตารางที่ 11 แสดงผลลัพธ์การจำแนกข้อมูลชุดที่ 1 ซึ่งเป็นชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพและเชิงปริมาณจำนวนเท่ากันและมีสัดส่วนความไม่สมดุลเท่ากับ 1:6.7 เมื่อพิจารณาค่า Pre Re และ F ของกลุ่ม No พบว่ามีค่าค่อนข้างสูงและทุกตัวแบบให้ค่าที่ใกล้เคียงกัน เมื่อพิจารณาค่า Pre Re และ F ของกลุ่ม Yes พบว่ามีค่าอยู่ในระดับกลางและทุกตัวแบบให้ค่าที่ใกล้เคียงกัน และเมื่อพิจารณาตัวแบบต่าง ๆ โดยใช้ค่า Acc ในการเปรียบเทียบ ผลที่ได้จากการเปรียบเทียบค่า Acc พบว่า ตัวแบบการจำแนกที่ให้ค่า Acc สูงสุดได้แก่ ANN RF LDA DT และ NB ตามลำดับ ซึ่งเมื่อพิจารณาจะพบว่าทุกตัวแบบจะค่าจำแนกคำตอบกลุ่ม No ได้ถูกต้องกว่ากลุ่ม Yes

ตารางที่ 12 แสดงผลลัพธ์การจำแนกภายใต้ชุดข้อมูลสายพันธุ์ข้าว (ตั้งต้น)

	Yes			No			Acc
	Pre	Re	F	Pre	Re	F	
LDA	0.923	**0.959**	**0.940**	**0.942**	0.893	**0.917**	**0.930**
NB	0.924	0.932	0.928	0.908	0.898	0.903	0.917
DT	**0.933**	0.933	0.933	0.912	**0.910**	0.910	0.923
RF	0.931	0.938	0.934	0.916	0.907	0.912	0.925
ANN (0.2,0.2,4)	0.929	0.948	0.939	0.929	0.903	0.916	0.929

ตารางที่ 12 แสดงผลลัพธ์การจำแนกข้อมูลชุดที่ 2 ซึ่งเป็นชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงปริมาณเท่านั้นและมีสัดส่วนความไม่สมดุลเท่ากับ 1:1.3 เมื่อพิจารณาค่า Pre Re และ F ของกลุ่ม No และ Yes พบว่ามีค่าค่อนข้างสูงและทุกตัวแบบให้ค่าที่ใกล้เคียงกัน ดังนั้นจึงสามารถพิจารณาตัวแบบต่าง ๆ โดยใช้ค่า Acc ในการเปรียบเทียบและผลที่ได้จากการเปรียบเทียบค่า Acc พบว่า ตัวแบบการจำแนกที่ให้ค่า Acc สูงที่สุดได้แก่ LDA ANN RF DT และ NB ตามลำดับ

ตารางที่ 13 แสดงผลลัพธ์การจำแนกภายใต้ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (ตั้งต้น)

	Yes			No			Acc
	Pre	Re	F	Pre	Re	F	
LDA	**0.400**	0.002	0.004	0.835	**1.000**	0.910	0.835
NB	0.247	**0.298**	**0.270**	**0.855**	0.820	0.837	0.733
DT	0.200	0.005	0.009	0.835	0.999	0.910	0.834
RF	0.353	0.005	0.009	0.835	0.999	0.910	0.834
ANN (0.2,0.2,17)	0.337	0.014	0.027	0.844	0.996	**0.914**	**0.842**

ตารางที่ 13 แสดงผลลัพธ์การจำแนกข้อมูลชุดที่ 3 ซึ่งเป็นชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพมากกว่าเชิงปริมาณและมีสัดส่วนความไม่สมดุลเท่ากับ 1:4.9 เมื่อพิจารณาค่า Pre Re และ F ของกลุ่ม No พบว่ามีค่าค่อนข้างสูงและทุกตัวแบบให้ค่าที่ใกล้เคียงกัน

เมื่อพิจารณาค่า Pre Re และ F ของกลุ่ม Yes พบว่ามีค่าค่อนข้างต่ำ นอกจากนี้ยังพบว่าตัวแบบ NB มีค่า Re และ F ของกลุ่ม Yes สูงที่สุดคือ 0.298 และ 0.270 ตามลำดับ ในขณะที่ตัวแบบ LDA DT RF และ ANN มีค่า Re และ F ของกลุ่ม Yes ค่อนข้างต่ำคือต่ำกว่า 0.014 และ 0.027 ตามลำดับ ซึ่งในที่นี้ไม่สามารถพิจารณาเปรียบเทียบตัวแบบการจำแนกที่มีประสิทธิภาพจากค่า Acc ได้

อีกทั้งยังพบว่าตัวแบบ NB มีค่า Acc ต่ำที่สุดแต่เมื่อพิจารณาเมทริกซ์ความสับสนพบว่าตัวแบบ NB (จากตารางผนวกที่ 12) มีค่า TP ที่สูงกว่าตัวแบบ LDA DT และ RF (จากตารางผนวกที่ 11 13 และ 14 ตามลำดับ) โดยตัวแบบ LDA DT และ RF เป็นตัวแบบที่มีค่า TP ต่ำมาก ซึ่งสามารถตีความได้ว่าตัวแบบเหล่านี้แทบจะจำแนกค่าคำตอบกลุ่ม Yes ไม่ถูกเลย ดังนั้นการเปรียบเทียบประสิทธิภาพเพียงใช้ค่า Acc ไม่เพียงพอต่อการประเมินประสิทธิภาพ ดังนั้นควรพิจารณาค่า Pre Re และ F เพื่อประกอบในการตัดสินใจด้วย

4.2 ผลการเปรียบเทียบประสิทธิภาพการจำแนกภายใต้ชุดข้อมูลสุ่มตัวอย่างแบบง่าย

ในหัวข้อนี้ผู้วิจัยจะนำเสนอผลการเปรียบเทียบประสิทธิภาพตัวแบบการจำแนกภายใต้ชุดข้อมูลสถาบันการเงิน (สุ่มตัวอย่างแบบง่าย) ชุดข้อมูลสายพันธุ์ข้าว (สุ่มตัวอย่างแบบง่าย) และชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (สุ่มตัวอย่างแบบง่าย) ดังตารางที่ 14 – 16 ตามลำดับ

ตารางที่ 14 แสดงผลลัพธ์การจำแนกภายใต้ชุดข้อมูลสถาบันการเงิน (สุ่มตัวอย่างแบบง่าย)

	Yes			No			Acc	Rank
	Pre	Re	F	Pre	Re	F		
LDA	**0.851**	0.843	0.847	0.851	**0.852**	0.848	0.847	4
NB	0.795	0.668	0.724	0.714	0.827	0.766	0.747	5
DT	0.846	0.911	0.877	0.904	0.834	0.868	0.873	2
RF	0.850	**0.927**	**0.887**	**0.920**	0.836	**0.876**	**0.882**	**1**
ANN (0.1,0.1,25)	0.838	0.908	0.871	0.900	0.824	0.860	0.866	3

ตารางที่ 14 แสดงผลลัพธ์การจำแนกข้อมูลชุดที่ 4 ซึ่งเป็นชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพและเชิงปริมาณจำนวนเท่ากันและมีสัดส่วนความไม่สมดุลเท่ากับ 1:1 เมื่อพิจารณาค่า Pre Re และ F ของกลุ่ม No และ Yes พบว่ามีค่าค่อนข้างสูงและทุกตัวแบบให้ค่าที่ใกล้เคียงกัน ดังนั้นจึงสามารถพิจารณาตัวแบบต่าง ๆ โดยใช้ค่า Acc ในการเปรียบเทียบและผลที่ได้จากการเปรียบเทียบค่า Acc พบว่า ตัวแบบการจำแนกที่ให้ค่า Acc สูงที่สุดได้แก่ RF DT ANN LDA และ NB ตามลำดับ

ตารางที่ 15 แสดงผลลัพธ์การจำแนกภายใต้ชุดข้อมูลสายพันธุ์ข้าว (สุ่มตัวอย่างแบบง่าย)

	Yes			No			Acc	Rank
	Pre	Re	F	Pre	Re	F		
LDA	0.929	**0.940**	**0.934**	**0.939**	0.928	**0.934**	**0.934**	**1**
NB	0.906	0.935	0.920	0.933	0.903	0.918	0.919	5
DT	0.932	0.928	0.930	0.928	0.932	0.930	0.930	3
RF	0.925	0.925	0.925	0.926	0.925	0.925	0.925	4
ANN (0.1,0.2,5)	**0.933**	0.926	0.929	0.924	**0.941**	0.930	0.933	2

ตารางที่ 15 แสดงผลลัพธ์การจำแนกข้อมูลชุดที่ 5 ซึ่งเป็นชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงปริมาณเท่านั้นและมีสัดส่วนความไม่สมดุลเท่ากับ 1:1 เมื่อพิจารณาค่า Pre Re และ F ของกลุ่ม No และ Yes พบว่ามีค่าค่อนข้างสูงและทุกตัวแบบให้ค่าที่ใกล้เคียงกัน ดังนั้นจึงสามารถพิจารณาตัวแบบต่าง ๆ โดยใช้ค่า Acc ในการเปรียบเทียบและผลที่ได้จากการเปรียบเทียบค่า Acc พบว่า ตัวแบบการจำแนกที่ให้ค่า Acc สูงที่สุดได้แก่ LDA ANN DT RF และ NB ตามลำดับ

ตารางที่ 16 แสดงผลลัพธ์การจำแนกภายใต้ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (สุ่มตัวอย่างแบบง่าย)

	Yes			No			Acc	Rank
	Pre	Re	F	Pre	Re	F		
LDA	0.609	0.673	**0.639**	**0.634**	0.567	0.599	0.620	2
NB	0.566	**0.716**	0.632	0.611	0.448	0.515	0.582	5
DT	0.607	0.624	0.616	0.614	0.596	0.605	0.610	4
RF	0.612	0.628	0.620	0.618	0.600	0.608	0.614	3
ANN (0.1,0.2,15)	**0.627**	0.616	0.617	0.625	**0.629**	**0.623**	**0.622**	**1**

ตารางที่ 16 แสดงผลลัพธ์การจำแนกข้อมูลชุดที่ 6 ซึ่งเป็นชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพมากกว่าเชิงปริมาณและมีสัดส่วนความไม่สมดุลเท่ากับ 1:1 เมื่อพิจารณาค่า Pre Re และ F ของกลุ่ม No และ Yes พบว่ามีค่าอยู่ในระดับกลางและทุกตัวแบบให้ค่าที่ใกล้เคียงกัน ดังนั้นจึงสามารถพิจารณาตัวแบบต่าง ๆ โดยใช้ค่า Acc ในการเปรียบเทียบและผลที่ได้จากการเปรียบเทียบค่า Acc พบว่า ตัวแบบการจำแนกที่ให้ค่า Acc สูงที่สุดได้แก่ ANN LDA RF DT และ NB ตามลำดับ

4.3 ผลการเปรียบเทียบประสิทธิภาพการจำแนกภายใต้ชุดข้อมูลแบ่งกลุ่มข้อมูลแบบเคมีน

ในหัวข้อนี้ผู้วิจัยจะนำเสนอผลการเปรียบเทียบประสิทธิภาพตัวแบบการจำแนกภายใต้ชุดข้อมูลสถาบันการเงิน (แบ่งกลุ่มข้อมูลแบบเคมีน) ชุดข้อมูลสายพันธุ์ข้าว (แบ่งกลุ่มข้อมูลแบบเคมีน) และชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (แบ่งกลุ่มข้อมูลแบบเคมีน) ดังตารางที่ 17

ตารางที่ 17 แสดงผลลัพธ์การจำแนกภายใต้ชุดข้อมูลสถาบันการเงิน (แบ่งกลุ่มข้อมูลแบบเคมีน)

	Yes			No			Acc	Rank
	Pre	Re	F	Pre	Re	F		
LDA	**0.847**	0.840	0.843	0.841	**0.848**	0.844	0.844	4
NB	0.796	0.662	0.721	0.711	0.828	0.764	0.745	5
DT	0.830	0.900	0.863	0.891	0.816	0.851	0.858	2
RF	0.839	**0.916**	**0.876**	**0.908**	0.824	**0.864**	**0.870**	**1**
ANN (0.2,0.1,22)	0.835	0.893	0.862	0.885	0.823	0.853	0.858	3

ตารางที่ 17 แสดงผลลัพธ์การจำแนกข้อมูลชุดที่ 7 ซึ่งเป็นชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพและเชิงปริมาณจำนวนเท่ากันและมีสัดส่วนความไม่สมดุลเท่ากับ 1:1 เมื่อพิจารณาค่า Pre Re และ F ของกลุ่ม No และ Yes พบว่ามีค่าค่อนข้างสูงและทุกตัวแบบให้ค่าที่ใกล้เคียงกัน ดังนั้นจึงสามารถพิจารณาตัวแบบต่าง ๆ โดยใช้ค่า Acc ในการเปรียบเทียบและผลที่ได้จากการเปรียบเทียบค่า Acc พบว่า ตัวแบบการจำแนกที่ให้ค่า Acc สูงที่สุดได้แก่ RF DT ANN LDA และ NB ตามลำดับ

ตารางที่ 18 แสดงผลลัพธ์การจำแนกภายใต้ชุดข้อมูลสายพันธุ์ข้าว (แบ่งกลุ่มข้อมูลแบบเคมีน)

	Yes			No			Acc	Rank
	Pre	Re	F	Pre	Re	F		
LDA	0.926	**0.936**	**0.931**	**0.935**	0.925	**0.930**	**0.930**	**1**
NB	0.902	0.933	0.917	0.930	0.899	0.914	0.916	5
DT	0.926	0.929	0.928	0.929	**0.926**	0.927	0.928	3
RF	0.915	0.928	0.922	0.927	0.914	0.921	0.921	4
ANN (0.1,0.2,6)	**0.933**	0.921	0.927	0.923	0.934	0.928	0.928	2

ตารางที่ 18 แสดงผลลัพธ์การจำแนกข้อมูลชุดที่ 8 ซึ่งเป็นชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงปริมาณเท่านั้นและมีสัดส่วนความไม่สมดุลเท่ากับ 1:1 เมื่อพิจารณาค่า Pre Re และ F ของกลุ่ม No และ Yes พบว่ามีค่าค่อนข้างสูงและทุกตัวแบบให้ค่าที่ใกล้เคียงกัน ดังนั้นจึงสามารถพิจารณาตัว

แบบต่าง ๆ โดยใช้ค่า Acc ในการเปรียบเทียบและผลที่ได้จากการเปรียบเทียบค่า Acc พบว่า ตัวแบบการจำแนกที่ให้ค่า Acc สูงที่สุดได้แก่ LDA ANN DT RF และ NB ตามลำดับ

ตารางที่ 19 แสดงผลลัพธ์การจำแนกภายใต้ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (แบ่งกลุ่มข้อมูลแบบเคมีน)

	Yes			No			Acc	Rank
	Pre	Re	F	Pre	Re	F		
LDA	0.610	0.678	0.642	0.638	0.567	**0.600**	0.622	2
NB	0.562	0.701	0.623	0.602	0.453	0.516	0.577	5
DT	0.589	0.599	0.594	0.593	0.582	0.587	0.591	4
RF	0.604	0.631	0.617	0.614	**0.587**	**0.600**	0.609	3
ANN (0.1,0.2,13)	**0.611**	**0.692**	**0.649**	**0.645**	0.559	0.598	**0.625**	**1**

ตารางที่ 19 แสดงผลลัพธ์การจำแนกข้อมูลชุดที่ 9 ซึ่งเป็นชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพมากกว่าเชิงปริมาณ และมีสัดส่วนความไม่สมดุลเท่ากับ 1:1 เมื่อพิจารณาค่า Pre Re และ F ของกลุ่ม No และ Yes พบว่ามีค่าอยู่ในระดับกลางและทุกตัวแบบให้ค่าที่ใกล้เคียงกัน ดังนั้นจึงสามารถพิจารณาตัวแบบต่าง ๆ โดยใช้ค่า Acc ในการเปรียบเทียบและผลที่ได้จากการเปรียบเทียบค่า Acc พบว่า ตัวแบบการจำแนกที่ให้ค่า Acc สูงที่สุดได้แก่ ANN LDA RF DT และ NB ตามลำดับ

4.4 ผลการเปรียบเทียบประสิทธิภาพการจำแนกภายใต้ชุดข้อมูลสุ่มตัวอย่างแบบง่ายกับชุดข้อมูลแบ่งกลุ่มข้อมูลแบบเคมีน

ในหัวข้อนี้ผู้วิจัยจะนำเสนอผลการเปรียบเทียบประสิทธิภาพตัวแบบการจำแนกภายใต้ชุดข้อมูลที่ปรับปรุงความสมดุลโดยวิธีการสุ่มตัวอย่างแบบง่ายและเทคนิคแบ่งกลุ่มข้อมูลแบบเคมีน ประกอบด้วย ชุดข้อมูลสถาบันการเงิน (สุ่มตัวอย่างแบบง่าย) เปรียบเทียบกับชุดข้อมูลสถาบันการเงิน (แบ่งกลุ่มข้อมูลแบบเคมีน) ชุดข้อมูลสายพันธุ์ข้าว (สุ่มตัวอย่างแบบง่าย) เปรียบเทียบกับชุดข้อมูลสายพันธุ์ข้าว (แบ่งกลุ่มข้อมูลแบบเคมีน) และชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (สุ่มตัวอย่างแบบง่าย) เปรียบเทียบกับชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (แบ่งกลุ่มข้อมูลแบบเคมีน) ดังตารางที่ 20

ตารางที่ 20 แสดงผลลัพธ์การจำแนกภายใต้ชุดข้อมูลสถาบันการเงิน (สุ่มตัวอย่างแบบง่าย) เปรียบเทียบกับชุดข้อมูลสถาบันการเงิน (แบ่งกลุ่มข้อมูลแบบเคมีน)

	ปรับปรุงความสมดุล	Yes			No			Acc
		Pre	Re	F	Pre	Re	F	
LDA	**แบบง่าย**	**0.851**	**0.843**	**0.847**	**0.851**	**0.852**	**0.848**	**0.847**
	เคมีน	0.847	0.840	0.843	0.841	0.848	0.844	0.844
NB	**แบบง่าย**	0.795	**0.668**	**0.724**	**0.714**	0.827	**0.766**	**0.747**
	เคมีน	**0.796**	0.662	0.721	0.711	**0.828**	0.764	0.745
DT	**แบบง่าย**	**0.846**	**0.911**	**0.877**	**0.904**	**0.834**	**0.868**	**0.873**
	เคมีน	0.830	0.900	0.863	0.891	0.816	0.851	0.858
RF	**แบบง่าย**	**0.850**	**0.927**	**0.887**	**0.920**	**0.836**	**0.876**	**0.882**
	เคมีน	0.839	0.916	0.876	0.908	0.824	0.864	0.870
ANN	**แบบง่าย**	**0.838**	**0.908**	**0.871**	**0.900**	**0.824**	**0.860**	**0.866**
	เคมีน	0.835	0.893	0.862	0.885	0.823	0.853	0.858

ตารางที่ 20 แสดงผลลัพธ์การจำแนกข้อมูลชุดที่ 4 และ 7 ซึ่งเป็นชุดข้อมูลที่ปรับปรุงความสมดุลด้วยการสุ่มตัวอย่างแบบง่ายและการแบ่งกลุ่มข้อมูลแบบเคมีน ตามลำดับ โดยทั้ง 2 ชุดเป็นชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพและเชิงปริมาณเท่ากันและมีสัดส่วนความไม่สมดุลเท่ากับ 1:1 โดยมีข้อสังเกตดังต่อไปนี้

เมื่อพิจารณาที่ตัวแบบ LDA DT RF และ ANN พบว่าการปรับปรุงชุดข้อมูลให้สมดุลด้วยการสุ่มตัวอย่างแบบง่ายให้ตัวแบบที่มีค่า Pre Re และ F ของกลุ่ม No และ Yes ที่สูงกว่าการแบ่งกลุ่มข้อมูลแบบเคมีน

เมื่อพิจารณาที่ตัวแบบ NB พบว่าการปรับปรุงชุดข้อมูลให้สมดุลด้วยการสุ่มตัวอย่างแบบง่ายให้ตัวแบบที่มีค่า Re และ F ของกลุ่ม Yes และค่า Pre และ F ของกลุ่ม No สูงกว่าการแบ่งกลุ่มข้อมูลแบบเคมีน

เมื่อพิจารณาที่ค่า Acc พบว่าการปรับปรุงชุดข้อมูลให้สมดุลด้วยการสุ่มตัวอย่างแบบง่ายให้ตัวแบบที่มีค่า Acc สูงกว่าการแบ่งกลุ่มข้อมูลแบบเคมีนทุกตัวแบบ

ตารางที่ 21 แสดงผลลัพธ์การจำแนกภายใต้ชุดข้อมูลสายพันธุ์ข้าว (สุ่มตัวอย่างแบบง่าย) เปรียบเทียบกับชุดข้อมูลสายพันธุ์ข้าว (แบ่งกลุ่มข้อมูลแบบเคมีน)

	ปรับปรุงความสมดุล	Yes			No			Acc
		Pre	Re	F	Pre	Re	F	
LDA	**แบบง่าย**	**0.929**	**0.940**	**0.934**	**0.939**	**0.928**	**0.934**	**0.934**
	เคมีน	0.926	0.936	0.931	0.935	0.925	0.930	0.930
NB	**แบบง่าย**	**0.906**	**0.935**	**0.920**	**0.933**	**0.903**	**0.918**	**0.919**
	เคมีน	0.902	0.933	0.917	0.930	0.899	0.914	0.916
DT	**แบบง่าย**	**0.932**	0.928	**0.930**	0.928	**0.932**	**0.930**	**0.930**
	เคมีน	0.926	**0.929**	0.928	**0.929**	0.926	0.927	0.928
RF	**แบบง่าย**	**0.925**	0.925	**0.925**	0.926	**0.925**	**0.925**	**0.925**
	เคมีน	0.915	**0.928**	0.922	**0.927**	0.914	0.921	0.921
ANN	**แบบง่าย**	**0.933**	**0.926**	**0.929**	**0.924**	**0.941**	**0.930**	**0.933**
	เคมีน	0.933	0.921	0.927	0.923	0.934	0.928	0.928

ตารางที่ 21 แสดงผลลัพธ์การจำแนกข้อมูลชุดที่ 5 และ 8 ซึ่งเป็นชุดข้อมูลที่ปรับปรุงชุดข้อมูลให้สมดุลด้วยการสุ่มตัวอย่างแบบง่ายและการแบ่งกลุ่มข้อมูลแบบเคมีน ตามลำดับ โดยทั้ง 2 ชุดเป็นชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงปริมาณเท่านั้นและมีสัดส่วนความไม่สมดุลเท่ากับ 1:1 โดยมีข้อสังเกตดังต่อไปนี้

เมื่อพิจารณาที่ตัวแบบ LDA NB และ ANN พบว่าการปรับปรุงชุดข้อมูลให้สมดุลด้วยการสุ่มตัวอย่างแบบง่ายให้ตัวแบบที่มีค่า Pre Re และ F ของกลุ่ม No และ Yes สูงกว่าการแบ่งกลุ่มข้อมูลแบบเคมีน

เมื่อพิจารณาที่ตัวแบบ DT และ RF พบว่าการปรับปรุงชุดข้อมูลให้สมดุลด้วยการสุ่มตัวอย่างแบบง่ายให้ตัวแบบที่มีค่า Pre และ F ของกลุ่ม Yes และค่า Re และ F ของกลุ่ม No สูงกว่าการแบ่งกลุ่มข้อมูลแบบเคมีน

เมื่อพิจารณาที่ค่า Acc พบว่าการปรับปรุงชุดข้อมูลให้สมดุลด้วยการสุ่มตัวอย่างแบบง่ายให้ตัวแบบที่มีค่า Acc สูงกว่าการแบ่งกลุ่มข้อมูลแบบเคมีนทุกตัวแบบ

ตารางที่ 22 แสดงผลลัพธ์การจำแนกภายใต้ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (สุ่มตัวอย่างแบบง่าย) เปรียบเทียบกับชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (แบ่งกลุ่มข้อมูลแบบเคมีน)

	ปรับปรุงความสมดุล	Yes			No			Acc
		Pre	Re	F	Pre	Re	F	
LDA	แบบง่าย	0.609	0.673	0.639	0.634	0.567	0.599	0.620
	เคมีน	**0.610**	**0.678**	**0.642**	**0.638**	**0.567**	**0.600**	**0.622**
NB	**แบบง่าย**	**0.566**	**0.716**	**0.632**	**0.611**	0.448	0.515	**0.582**
	เคมีน	0.562	0.701	0.623	0.602	**0.453**	**0.516**	0.577
DT	**แบบง่าย**	**0.607**	**0.624**	**0.616**	**0.614**	**0.596**	**0.605**	**0.610**
	เคมีน	0.589	0.599	0.594	0.593	0.582	0.587	0.591
RF	**แบบง่าย**	**0.612**	0.628	**0.620**	**0.618**	**0.600**	**0.608**	**0.614**
	เคมีน	0.604	**0.631**	0.617	0.614	0.587	0.600	0.609
ANN	แบบง่าย	0.627	0.616	0.617	0.625	0.629	0.623	0.622
	เคมีน	**0.611**	**0.692**	**0.649**	**0.645**	**0.559**	**0.598**	**0.625**

ตารางที่ 22 แสดงผลลัพธ์การจำแนกข้อมูลชุดที่ 6 และ 9 ซึ่งเป็นชุดข้อมูลที่ปรับปรุงความสมดุลด้วยการสุ่มตัวอย่างแบบง่ายและการแบ่งกลุ่มข้อมูลแบบเคมีน ตามลำดับ โดยทั้ง 2 ชุดเป็นชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพมากกว่าเชิงปริมาณและมีสัดส่วนความไม่สมดุลเท่ากับ 1:1 โดยมีข้อสังเกตดังต่อไปนี้

เมื่อพิจารณาที่ตัวแบบ LDA และ ANN พบว่าการปรับปรุงชุดข้อมูลให้สมดุลด้วยการแบ่งกลุ่มข้อมูลแบบเคมีนให้ตัวแบบที่มีค่า Pre Re และ F ของกลุ่ม No และ Yes สูงกว่าการสุ่มตัวอย่างแบบง่าย

เมื่อพิจารณาที่ตัวแบบ NB พบว่าการปรับปรุงชุดข้อมูลให้สมดุลด้วยการสุ่มตัวอย่างแบบง่ายให้ตัวแบบที่มีค่า Pre Re และ F ของกลุ่ม Yes และค่า Pre ของกลุ่ม No สูงกว่าการแบ่งกลุ่มข้อมูลแบบเคมีน

เมื่อพิจารณาที่ตัวแบบ DT พบว่าการปรับปรุงชุดข้อมูลให้สมดุลด้วยการสุ่มตัวอย่างแบบง่ายให้ตัวแบบที่มีค่า Pre Re และ F ของกลุ่ม No และ Yes สูงกว่าแบ่งกลุ่มข้อมูลแบบเคมีน

เมื่อพิจารณาที่ตัวแบบ RF พบว่าการปรับปรุงชุดข้อมูลให้สมดุลด้วยการสุ่มตัวอย่างแบบง่ายให้ตัวแบบที่มีค่า Pre และ F ของกลุ่ม Yes และค่า Pre Re และ F ของกลุ่ม No สูงกว่าการแบ่งกลุ่มข้อมูลแบบเคมีน

เมื่อพิจารณาที่ค่า Acc พบว่าการปรับปรุงชุดข้อมูลให้สมดุลด้วยการสุ่มตัวอย่างแบบง่ายให้ตัวแบบที่มีค่า Acc สูงกว่าเทคนิคการแบ่งกลุ่มข้อมูลแบบเคมีนได้แก่ตัวแบบ NB DT และ RF

บทที่ 5
สรุปผลการวิจัย

การวิจัยเรื่อง การเปรียบเทียบเทคนิคการเรียนรู้ของเครื่องเพื่อสร้างตัวแบบการจำแนกด้วยการปรับปรุงชุดข้อมูลอสมดุล มีวัตถุประสงค์เพื่อศึกษาประสิทธิภาพเทคนิคการจำแนกภายใต้ชุดข้อมูล 3 ชุด ที่มีจำนวนของตัวแปรอิสระเชิงคุณภาพและเชิงปริมาณที่แตกต่างกันและปรับปรุงชุดข้อมูลอสมดุลให้สมดุลด้วยวิธีสุ่มลดได้แก่ การสุ่มตัวอย่างแบบง่ายและการแบ่งกลุ่มข้อมูลแบบเคมีน โดยข้อมูลที่ใช้ในการศึกษามีจำนวน 3 ชุด ประกอบด้วย ชุดข้อมูลสถาบันการเงิน ชุดข้อมูลสายพันธุ์ข้าวและชุดข้อมูลนักวิทยาศาสตร์ข้อมูล ในขั้นตอนของการพัฒนาตัวแบบการจำแนกจะแบ่งชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบด้วยหลักการ 5-Fold จากนั้นนำชุดข้อมูลมาใช้ในการสร้างตัวแบบการจำแนกโดยใช้เทคนิคการวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์ เทคนิคนาอีฟเบย์ ต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 เทคนิคป่าสุ่มและโครงข่ายประสาทเทียม จากนั้นทำการเปรียบเทียบประสิทธิภาพโดยใช้ค่าความแม่น ค่าความเที่ยง ค่าการเรียกคืนและค่าประสิทธิภาพ

ผลที่ได้จากการศึกษาสามารถสรุปได้ดังนี้

1. จากชุดข้อมูลสถาบันการเงินที่ปรับปรุงชุดข้อมูลอสมดุลให้สมดุลด้วยการสุ่มตัวอย่างแบบง่ายและการแบ่งกลุ่มข้อมูลแบบเคมีน ซึ่งเป็นชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพและปริมาณเท่ากันพบว่า เทคนิคป่าสุ่มเป็นเทคนิคการจำแนกที่มีประสิทธิภาพมากที่สุด (จากผลลัพธ์ในตารางที่ 15 และ ตารางที่ 17) กล่าวคือ เทคนิคป่าสุ่มทำงานได้ดีภายใต้ชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพและปริมาณเท่ากัน

2. จากชุดข้อมูลสายพันธุ์ข้าวที่ปรับปรุงชุดข้อมูลอสมดุลให้สมดุลด้วยการสุ่มตัวอย่างแบบง่ายและการแบ่งกลุ่มข้อมูลแบบเคมีน ซึ่งเป็นชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงปริมาณเท่านั้นพบว่าการวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์เป็นเทคนิคการจำแนกที่มีประสิทธิภาพมากที่สุด (จากผลลัพธ์ในตารางที่ 15 และตารางที่ 18) กล่าวคือการวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์ทำงานได้ดีภายใต้ชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงปริมาณทุกตัว

3. จากชุดข้อมูลนักวิทยาศาสตร์ข้อมูลที่ปรับปรุงชุดข้อมูลอสมดุลให้สมดุลด้วยการสุ่มตัวอย่างแบบง่ายและการแบ่งกลุ่มข้อมูลแบบเคมีน ซึ่งเป็นชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพมากกว่าเชิงปริมาณพบว่า โครงข่ายประสาทเทียมเป็นเทคนิคการจำแนกที่มีประสิทธิภาพมากที่สุด (จากผลลัพธ์ในตารางที่ 16 และตารางที่ 19) กล่าวคือโครงข่ายประสาทเทียมทำงานได้ดีภายใต้ชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพมากกว่าปริมาณ

4. การปรับปรุงชุดข้อมูลอสมดุลให้สมดุลด้วยการสุ่มตัวอย่างแบบง่ายให้ตัวแบบการจำแนกที่มีประสิทธิภาพสูงการแบ่งกลุ่มข้อมูลแบบเคมีน อีกทั้งวิธีการสุ่มตัวอย่างแบบง่ายเป็นวิธีที่ง่ายและซับซ้อนน้อยกว่าการแบ่งกลุ่มข้อมูลแบบเคมีน

5. การประเมินประสิทธิภาพตัวแบบการจำแนกที่สร้างจากชุดข้อมูลอสมดุล โดยใช้เพียงค่าความแม่นอย่างเดียวอาจไม่เพียงพอต่อการประเมิน ดังนั้นควรนำค่าความเที่ยง ค่าการเรียกคืนและค่าประสิทธิภาพเพื่อมาประกอบในการตัดสินใจเลือกตัวแบบการจำแนกที่มีประสิทธิภาพด้วยเพื่อนำมาใช้งาน

5.1 ข้อเสนอแนะ

จากการวิจัยในครั้งนี้ เพื่อเป็นแนวทางในการพัฒนางานวิจัยให้มีประสิทธิภาพมากขึ้น ผู้วิจัยมีข้อเสนอแนะดังนี้

1. การนำเทคนิคการปรับปรุงชุดข้อมูลอสมดุลให้สมดุลด้วยเทคนิคอื่น ๆ มาประยุกต์ใช้ เช่น การสุ่มเพิ่ม เป็นต้น

2. วิธีการแบ่งชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบด้วยเทคนิคอื่น ๆ เช่น 10 Fold เป็นต้น

3. การประยุกต์ใช้เทคนิคการจำแนกอื่น ๆ เช่น เทคนิคเพื่อนบ้านใกล้ที่สุด k ตัว และการถดถอยลอจิสติก เป็นต้น

4. การใช้วิธีการเลือกค่า k ที่เหมาะสมในเทคนิคการแบ่งกลุ่มข้อมูลแบบเคมีนด้วยวิธีอื่น ๆ

5. การเปลี่ยนพารามิเตอร์ได้แก่ ค่าอัตราการเรียนรู้ ค่าโมเมนตัม จำนวนชั้นซ่อนและจำนวนโหนดในชั้นซ่อนด้วยค่าอื่น ๆ

6. คัดเลือกชุดข้อมูลที่มีลักษณะแตกต่างกันจำนวนหลาย ๆ ชุดเพื่อสามารถเปรียบเทียบประสิทธิภาพการจำแนกได้ชัดเจนยิ่งขึ้น

7. เลือกชุดข้อมูลที่มีสัดส่วนของจำนวนตัวแปรอิสระเชิงคุณภาพและเชิงปริมาณ เพื่อศึกษาความสัมพันธ์ของเทคนิคการจำแนกกับลักษณะของตัวแปรอิสระ

ภาคผนวก ก
เมทริกซ์ความสับสนของตัวแบบการจำแนก

ชุดข้อมูลสถาบันการเงิน (ตั้งต้น)

ตารางผนวกที่ 1 เมทริกซ์ความสับสนของตัวแบบการวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์ ภายใต้ชุดข้อมูลสถาบันการเงิน (ตั้งต้น)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	396	222	5103	375	0.641	0.514	0.570	0.932	0.958	0.945	0.902
2	387	237	5089	385	0.620	0.501	0.554	0.930	0.956	0.942	0.898
3	390	229	5097	382	0.630	0.505	0.561	0.930	0.957	0.943	0.900
4	382	255	5071	390	0.600	0.495	0.542	0.929	0.952	0.940	0.894
5	387	224	5102	385	0.633	0.501	0.560	0.930	0.958	0.944	0.900
Average					0.625	0.503	0.557	0.930	0.956	0.943	0.899

ตารางผนวกที่ 2 เมทริกซ์ความสับสนของตัวแบบนาอีฟเบย์ ภายใต้ชุดข้อมูลสถาบันการเงิน (ตั้งต้น)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	476	686	4639	295	0.410	0.617	0.492	0.940	0.871	0.904	0.839
2	461	679	4647	311	0.404	0.597	0.482	0.937	0.873	0.904	0.838
3	448	711	4615	324	0.387	0.580	0.464	0.934	0.867	0.899	0.830
4	440	615	4711	332	0.417	0.570	0.482	0.934	0.885	0.909	0.845
5	444	585	4741	328	0.431	0.575	0.493	0.935	0.890	0.912	0.850
Average					0.410	0.588	0.483	0.936	0.877	0.906	0.840

ตารางผนวกที่ 3 เมทริกซ์ความสับสนของตัวแบบต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 ภายใต้ชุดข้อมูลสถาบันการเงิน (ตั้งต้น)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	407	257	5068	364	0.613	0.528	0.567	0.933	0.952	0.942	0.898
2	411	339	4987	361	0.548	0.532	0.540	0.932	0.936	0.934	0.885
3	392	274	5052	380	0.589	0.508	0.545	0.930	0.949	0.939	0.893
4	394	268	5058	378	0.595	0.510	0.550	0.930	0.950	0.940	0.894

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
5	407	267	5059	365	0.604	0.527	0.563	0.933	0.950	0.941	0.896
Average					0.589	0.522	0.554	0.932	0.947	0.939	0.893

ตารางผนวกที่ 4 เมทริกซ์ความสับสนของตัวแบบเทคนิคป่าสุ่ม ภายใต้ชุดข้อมูลสถาบันการเงิน (ตั้งต้น)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	397	187	5138	374	0.680	0.515	0.586	0.932	0.965	0.948	0.908
2	376	199	5127	396	0.654	0.487	0.558	0.928	0.963	0.945	0.902
3	363	182	5144	409	0.666	0.470	0.551	0.926	0.966	0.946	0.903
4	349	184	5142	423	0.655	0.452	0.535	0.924	0.965	0.944	0.900
5	366	181	5145	406	0.669	0.474	0.555	0.927	0.966	0.946	0.904
Average					0.665	0.480	0.557	0.927	0.965	0.946	0.903

ตารางผนวกที่ 5 แสดงผลลัพธ์การจำแนกโครงข่ายประสาทเทียม ภายใต้ชุดข้อมูลสถาบันการเงิน (ตั้งต้น)

(l, m, h)	Yes			No			Acc
	Pre	Re	F	Pre	Re	F	
(0.1, 0.1, 21)	0.658	0.517	0.577	0.932	0.961	0.946	0.904
(0.1, 0.1, 22)	0.680	0.462	0.548	0.926	0.968	0.946	0.904
(0.1, 0.1, 23)	0.679	0.441	0.531	0.923	0.970	0.946	0.903
(0.1, 0.1, 24)	0.680	0.450	0.541	0.924	0.969	0.946	0.904
(0.1, 0.1, 25)	0.698	0.381	0.492	0.916	0.976	0.945	0.901
(0.2,0.1,21)	**0.654**	**0.520**	**0.577**	**0.933**	**0.960**	**0.946**	**0.904**
(0.2, 0.1, 22)	0.676	0.464	0.549	0.926	0.967	0.946	0.904
(0.2, 0.1, 23)	0.680	0.443	0.532	0.923	0.969	0.946	0.903
(0.2, 0.1, 24)	0.673	0.469	0.553	0.926	0.967	0.946	0.904
(0.2, 0.1, 25)	0.707	0.369	0.485	0.914	0.978	0.945	0.901
(0.1, 0.2, 21)	0.626	0.580	0.602	0.940	0.950	0.945	0.903
(0.1, 0.2, 22)	0.670	0.467	0.550	0.926	0.967	0.946	0.903

(l, m, h)	Yes			No			Acc
	Pre	Re	F	Pre	Re	F	
(0.1, 0.2, 23)	0.651	0.517	0.576	0.932	0.960	0.946	0.904
(0.1, 0.2, 24)	0.660	0.483	0.557	0.928	0.964	0.946	0.903
(0.1, 0.2, 25)	0.695	0.398	0.505	0.918	0.975	0.945	0.902
(0.2, 0.2, 21)	0.629	0.569	0.597	0.938	0.951	0.945	0.903
(0.2, 0.2, 22)	0.671	0.458	0.544	0.925	0.967	0.946	0.903
(0.2, 0.2, 23)	0.655	0.511	0.574	0.931	0.961	0.946	0.904
(0.2, 0.2, 24)	0.662	0.478	0.555	0.927	0.965	0.946	0.903
(0.2, 0.2, 25)	0.700	0.396	0.506	0.918	0.975	0.946	0.902

ชุดข้อมูลสายพันธุ์ข้าว (ตั้งต้น)

ตารางผนวกที่ 6 เมทริกซ์ความสับสนของตัวแบบการวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์ ภายใต้ชุดข้อมูลสายพันธุ์ข้าว (ตั้งต้น)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	418	35	291	18	0.923	0.959	0.940	0.942	0.893	0.917	0.930
2	414	35	291	22	0.922	0.950	0.936	0.930	0.893	0.911	0.925
3	418	34	292	18	0.925	0.959	0.941	0.942	0.896	0.918	0.932
4	414	29	297	22	0.935	0.950	0.942	0.931	0.911	0.921	0.933
5	417	33	293	19	0.927	0.956	0.941	0.939	0.899	0.918	0.932
Average					0.926	0.955	0.940	0.937	0.898	0.917	0.930

ตารางผนวกที่ 7 เมทริกซ์ความสับสนของตัวแบบนาอีฟเบย์ ภายใต้ชุดข้อมูลสายพันธุ์ข้าว (ตั้งต้น)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	406	32	294	30	0.927	0.931	0.929	0.907	0.902	0.905	0.919
2	404	33	293	32	0.924	0.927	0.926	0.902	0.899	0.900	0.915
3	410	38	288	26	0.915	0.940	0.928	0.917	0.883	0.900	0.916
4	402	29	297	34	0.933	0.922	0.927	0.897	0.911	0.904	0.917

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
5	410	35	291	26	0.921	0.940	0.931	0.918	0.893	0.905	0.920
Average					0.924	0.932	0.928	0.908	0.898	0.903	0.917

ตารางผนวกที่ 8 เมทริกซ์ความสับสนของตัวแบบต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 ภายใต้ชุดข้อมูลสายพันธุ์ข้าว (ตั้งต้น)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	401	23	303	35	0.946	0.920	0.933	0.896	0.929	0.913	0.924
2	406	31	295	30	0.929	0.931	0.930	0.908	0.905	0.906	0.920
3	417	42	284	19	0.908	0.956	0.932	0.937	0.871	0.903	0.920
4	397	23	303	39	0.945	0.911	0.928	0.886	0.929	0.907	0.919
5	414	28	298	22	0.937	0.950	0.943	0.931	0.914	0.923	0.934
Average					0.933	0.933	0.933	0.912	0.910	0.910	0.923

ตารางผนวกที่ 9 เมทริกซ์ความสับสนของตัวแบบเทคนิคป่าสุ่ม ภายใต้ชุดข้อมูลสายพันธุ์ข้าว (ตั้งต้น)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	407	26	300	29	0.940	0.933	0.937	0.912	0.920	0.916	0.928
2	411	32	294	25	0.928	0.943	0.935	0.922	0.902	0.912	0.925
3	410	40	286	26	0.911	0.940	0.926	0.917	0.877	0.897	0.913
4	402	26	300	34	0.939	0.922	0.931	0.898	0.920	0.909	0.921
5	414	27	299	22	0.939	0.950	0.944	0.931	0.917	0.924	0.936
Average					0.931	0.938	0.934	0.916	0.907	0.912	0.925

ตารางผนวกที่ 10 แสดงผลลัพธ์การจำแนกโครงข่ายประสาทเทียม ภายใต้ชุดข้อมูลสายพันธุ์ข้าว (ตั้งต้น)

(l, m, h)	Yes			No			Acc
	Pre	Re	F	Pre	Re	F	
(0.1, 0.1, 2)	0.946	0.926	0.936	0.904	0.929	0.916	0.927
(0.1, 0.1, 3)	0.942	0.930	0.936	0.909	0.923	0.916	0.927
(0.1, 0.1, 4)	0.933	0.939	0.936	0.917	0.910	0.913	0.926
(0.1, 0.1, 5)	0.944	0.928	0.936	0.907	0.926	0.916	0.927
(0.1, 0.1, 6)	0.941	0.931	0.936	0.909	0.922	0.915	0.927
(0.2, 0.1, 2)	0.945	0.926	0.935	0.904	0.928	0.916	0.927
(0.2, 0.1, 3)	0.943	0.930	0.936	0.909	0.924	0.916	0.928
(0.2, 0.1, 4)	0.932	0.939	0.936	0.918	0.909	0.913	0.926
(0.2, 0.1, 5)	0.944	0.928	0.936	0.906	0.926	0.916	0.927
(0.2, 0.1, 6)	0.941	0.930	0.935	0.909	0.921	0.915	0.927
(0.1, 0.2, 2)	0.945	0.926	0.936	0.904	0.928	0.916	0.927
(0.1, 0.2, 3)	0.946	0.925	0.936	0.903	0.929	0.916	0.927
(0.1, 0.2, 4)	0.929	0.949	0.939	0.930	0.903	0.916	0.929
(0.1, 0.2, 5)	0.946	0.923	0.934	0.901	0.929	0.915	0.926
(0.1, 0.2, 6)	0.948	0.920	0.934	0.898	0.933	0.915	0.925
(0.2, 0.2, 2)	0.945	0.926	0.935	0.904	0.928	0.916	0.927
(0.2, 0.2, 3)	0.945	0.925	0.935	0.903	0.928	0.915	0.927
(0.2, 0.2, 4)	**0.929**	**0.948**	**0.939**	**0.929**	**0.903**	**0.916**	**0.929**
(0.2, 0.2, 5)	0.947	0.922	0.934	0.900	0.931	0.915	0.926
(0.2, 0.2, 6)	0.948	0.921	0.934	0.898	0.932	0.915	0.925

ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (ตั้งต้น)

ตารางผนวกที่ 11 เมทริกซ์ความสับสนของตัวแบบการวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์ ภายใต้ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (ตั้งต้น)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	1	0	1494	295	1.000	0.003	0.007	0.835	1.000	0.910	0.835
2	0	1	1493	296	0.000	0.000	0.000	0.835	0.999	0.910	0.834
3	0	0	1494	297	0.000	0.000	0.000	0.834	1.000	0.910	0.834
4	0	0	1495	297	0.000	0.000	0.000	0.834	1.000	0.910	0.834
5	2	0	1495	295	1.000	0.007	0.013	0.835	1.000	0.910	0.835
Average					0.400	0.002	0.004	0.835	1.000	0.910	0.835

ตารางผนวกที่ 12 เมทริกซ์ความสับสนของตัวแบบนาอีฟเบย์ ภายใต้ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (ตั้งต้น)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	76	262	1232	220	0.225	0.257	0.240	0.848	0.825	0.836	0.731
2	86	258	1236	210	0.250	0.291	0.269	0.855	0.827	0.841	0.739
3	105	307	1187	192	0.255	0.354	0.296	0.861	0.795	0.826	0.721
4	95	254	1241	202	0.272	0.320	0.294	0.860	0.830	0.845	0.746
5	80	266	1229	217	0.231	0.269	0.249	0.850	0.822	0.836	0.730
Average					0.247	0.298	0.270	0.855	0.820	0.837	0.733

ตารางผนวกที่ 13 เมทริกซ์ความสับสนของตัวแบบต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 ภายใต้ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (ตั้งต้น)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	4	3	1491	292	0.571	0.014	0.026	0.836	0.998	0.910	0.835
2	3	4	1490	293	0.429	0.010	0.020	0.836	0.997	0.909	0.834
3	0	0	1494	297	0.000	0.000	0.000	0.834	1.000	0.910	0.834

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
4	0	0	1495	297	0.000	0.000	0.000	0.834	1.000	0.910	0.834
5	0	0	1495	297	0.000	0.000	0.000	0.834	1.000	0.910	0.834
Average					0.200	0.005	0.009	0.835	0.999	0.910	0.834

ตารางผนวกที่ 14 เมทริกซ์ความสับสนของตัวแบบเทคนิคป่าสุ่ม ภายใต้ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (ตั้งต้น)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	0	3	1491	296	0.000	0.000	0.000	0.834	0.998	0.909	0.833
2	0	1	1493	296	0.000	0.000	0.000	0.835	0.999	0.910	0.834
3	2	1	1493	295	0.667	0.007	0.013	0.835	0.999	0.910	0.835
4	3	2	1493	294	0.600	0.010	0.020	0.835	0.999	0.910	0.835
5	2	2	1493	295	0.500	0.007	0.013	0.835	0.999	0.910	0.834
Average					0.353	0.005	0.009	0.835	0.999	0.910	0.834

ตารางผนวกที่ 15 แสดงผลลัพธ์การจำแนกโครงข่ายประสาทเทียม ภายใต้ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (ตั้งต้น)

(l, m, h)	Yes			No			Acc
	Pre	Re	F	Pre	Re	F	
(0.1, 0.1, 13)	0.000	0.000	0.000	0.837	1.000	0.911	0.837
(0.1, 0.1, 14)	0.000	0.001	0.000	0.844	1.000	0.915	0.844
(0.1, 0.1, 15)	0.167	0.003	0.001	0.835	0.999	0.909	0.834
(0.1, 0.1, 16)	0.275	0.003	0.005	0.835	0.999	0.910	0.834
(0.1, 0.1, 17)	0.200	0.002	0.003	0.835	0.999	0.910	0.834
(0.2, 0.1, 13)	0.000	0.000	0.000	0.834	0.999	0.909	0.833
(0.2, 0.1, 14)	0.000	0.001	0.000	0.834	1.000	0.910	0.834
(0.2, 0.1, 15)	0.167	0.003	0.001	0.835	0.999	0.909	0.834
(0.2, 0.1, 16)	0.300	0.003	0.005	0.835	0.999	0.910	0.834

(l, m, h)	Yes			No			Acc
	Pre	Re	F	Pre	Re	F	
(0.2, 0.1, 17)	0.433	0.003	0.005	0.835	0.999	0.910	0.834
(0.1, 0.2, 13)	0.000	0.001	0.000	0.668	1.000	0.728	0.834
(0.1, 0.2, 14)	0.050	0.001	0.001	0.834	0.999	0.909	0.834
(0.1, 0.2, 15)	0.280	0.009	0.018	0.835	0.997	0.909	0.834
(0.1, 0.2, 16)	0.214	0.005	0.009	0.835	0.999	0.910	0.834
(0.1, 0.2, 17)	0.289	0.008	0.016	0.835	0.997	0.909	0.833
(0.2, 0.2, 13)	0.000	0.001	0.000	0.834	1.000	0.910	0.835
(0.2, 0.2, 14)	0.057	0.001	0.003	0.834	0.999	0.909	0.834
(0.2, 0.2, 15)	0.367	0.010	0.019	0.835	0.997	0.909	0.834
(0.2, 0.2, 16)	0.177	0.005	0.010	0.835	0.998	0.909	0.834
(0.2,0.2,17)	**0.337**	**0.014**	**0.027**	**0.844**	**0.996**	**0.914**	**0.842**

ชุดข้อมูลสถาบันการเงิน (สุ่มตัวอย่างแบบง่าย)

ตารางผนวกที่ 16 เมทริกซ์ความสับสนของตัวแบบการวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์ ภายใต้ชุดข้อมูลสถาบันการเงิน (สุ่มตัวอย่างแบบง่าย)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	666	123	648	105	0.844	0.864	0.854	0.861	0.840	0.850	0.852
2	646	116	656	126	0.848	0.837	0.842	0.839	0.850	0.844	0.843
3	661	104	668	111	0.864	0.856	0.860	0.858	0.865	0.861	0.861
4	631	107	665	141	0.855	0.817	0.836	0.825	0.861	0.843	0.839
5	648	121	651	124	0.843	0.839	0.841	0.840	0.843	0.842	0.841
Average					0.851	0.843	0.847	0.851	0.852	0.848	0.847

ตารางผนวกที่ 17 เมทริกซ์ความสับสนของตัวแบบนาอีฟเบย์ ภายใต้ชุดข้อมูลสถาบันการเงิน (สุ่มตัวอย่างแบบง่าย)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	449	101	670	322	0.816	0.582	0.680	0.675	0.869	0.760	0.726
2	546	152	620	226	0.782	0.707	0.743	0.733	0.803	0.766	0.755
3	531	130	642	241	0.803	0.688	0.741	0.727	0.832	0.776	0.760
4	525	135	637	247	0.795	0.680	0.733	0.721	0.825	0.769	0.753
5	525	151	621	247	0.777	0.680	0.725	0.715	0.804	0.757	0.742
Average					0.795	0.668	0.724	0.714	0.827	0.766	0.747

ตารางผนวกที่ 18 เมทริกซ์ความสับสนของตัวแบบต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 ภายใต้ชุดข้อมูลสถาบันการเงิน (สุ่มตัวอย่างแบบง่าย)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	705	137	634	66	0.837	0.914	0.874	0.906	0.822	0.862	0.868
2	700	122	650	72	0.852	0.907	0.878	0.900	0.842	0.870	0.874
3	721	121	651	51	0.856	0.934	0.893	0.927	0.843	0.883	0.889
4	692	138	634	80	0.834	0.896	0.864	0.888	0.821	0.853	0.859
5	698	122	650	74	0.851	0.904	0.877	0.898	0.842	0.869	0.873
Average					0.846	0.911	0.877	0.904	0.834	0.868	0.873

ตารางผนวกที่ 19 เมทริกซ์ความสับสนของตัวแบบเทคนิคป่าสุ่ม ภายใต้ชุดข้อมูลสถาบันการเงิน (สุ่มตัวอย่างแบบง่าย)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	724	143	628	47	0.835	0.939	0.884	0.930	0.815	0.869	0.877
2	714	117	655	58	0.859	0.925	0.891	0.919	0.848	0.882	0.887
3	719	122	650	53	0.855	0.931	0.892	0.925	0.842	0.881	0.887
4	703	118	654	69	0.856	0.911	0.883	0.905	0.847	0.875	0.879

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
5	716	131	641	56	0.845	0.927	0.884	0.920	0.830	0.873	0.879
Average					0.850	0.927	0.887	0.920	0.836	0.876	0.882

ตารางผนวกที่ 20 แสดงผลลัพธ์การจำแนกโครงข่ายประสาทเทียม ภายใต้ชุดข้อมูลสถาบันการเงิน (สุ่มตัวอย่างแบบง่าย)

(l, m, h)	Yes			No			Acc
	Pre	Re	F	Pre	Re	F	
(0.1, 0.1, 21)	0.826	0.925	0.873	0.916	0.804	0.856	0.865
(0.1, 0.1, 22)	0.853	0.870	0.861	0.869	0.849	0.858	0.860
(0.1, 0.1, 23)	0.856	0.864	0.860	0.864	0.854	0.858	0.859
(0.1, 0.1, 24)	0.849	0.879	0.864	0.875	0.844	0.859	0.861
(0.1,0.1,25)	**0.838**	**0.908**	**0.871**	**0.900**	**0.824**	**0.860**	**0.866**
(0.2, 0.1, 21)	0.827	0.925	0.873	0.916	0.805	0.857	0.865
(0.2, 0.1, 22)	0.853	0.870	0.861	0.868	0.849	0.858	0.860
(0.2, 0.1, 23)	0.853	0.870	0.861	0.868	0.850	0.858	0.860
(0.2, 0.1, 24)	0.849	0.878	0.863	0.874	0.844	0.859	0.861
(0.2, 0.1, 25)	0.838	0.906	0.870	0.898	0.824	0.859	0.865
(0.1, 0.2, 21)	0.826	0.922	0.871	0.912	0.805	0.855	0.863
(0.1, 0.2, 22)	0.852	0.868	0.859	0.867	0.848	0.857	0.858
(0.1, 0.2, 23)	0.850	0.870	0.860	0.868	0.846	0.857	0.858
(0.1, 0.2, 24)	0.852	0.870	0.861	0.868	0.848	0.858	0.860
(0.1, 0.2, 25)	0.835	0.903	0.867	0.895	0.821	0.856	0.862
(0.2, 0.2, 21)	0.825	0.923	0.871	0.913	0.803	0.854	0.863
(0.2, 0.2, 22)	0.851	0.867	0.859	0.866	0.848	0.856	0.857
(0.2, 0.2, 23)	0.852	0.873	0.862	0.872	0.848	0.859	0.860
(0.2, 0.2, 24)	0.850	0.880	0.864	0.875	0.843	0.859	0.862
(0.2, 0.2, 25)	0.831	0.917	0.872	0.907	0.813	0.857	0.865

ชุดข้อมูลสายพันธุ์ข้าว (สุ่มตัวอย่างแบบง่าย)

ตารางผนวกที่ 21 เมทริกซ์ความสับสนของตัวแบบการวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์ ภายใต้ชุดข้อมูลสายพันธุ์ข้าว (สุ่มตัวอย่างแบบง่าย)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	306	23	303	20	0.930	0.939	0.934	0.938	0.929	0.934	0.934
2	307	22	304	19	0.933	0.942	0.937	0.941	0.933	0.937	0.937
3	310	29	297	16	0.914	0.951	0.932	0.949	0.911	0.930	0.931
4	302	18	308	24	0.944	0.926	0.935	0.928	0.945	0.936	0.936
5	307	25	301	19	0.925	0.942	0.933	0.941	0.923	0.932	0.933
Average					0.929	0.940	0.934	0.939	0.928	0.934	0.934

ตารางผนวกที่ 22 เมทริกซ์ความสับสนของตัวแบบเทคนิคนาอีฟเบย์ ภายใต้ชุดข้อมูลสายพันธุ์ข้าว (สุ่มตัวอย่างแบบง่าย)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	306	26	300	20	0.922	0.939	0.930	0.938	0.920	0.929	0.929
2	305	33	293	21	0.902	0.936	0.919	0.933	0.899	0.916	0.917
3	307	39	287	19	0.887	0.942	0.914	0.938	0.880	0.908	0.911
4	301	31	295	25	0.907	0.923	0.915	0.922	0.905	0.913	0.914
5	305	29	297	21	0.913	0.936	0.924	0.934	0.911	0.922	0.923
Average					0.906	0.935	0.920	0.933	0.903	0.918	0.919

ตารางผนวกที่ 23 เมทริกซ์ความสับสนของตัวแบบต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 ภายใต้ชุดข้อมูลสายพันธุ์ข้าว (สุ่มตัวอย่างแบบง่าย)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	296	18	308	30	0.943	0.908	0.925	0.911	0.945	0.928	0.926
2	303	19	307	23	0.941	0.929	0.935	0.930	0.942	0.936	0.936
3	309	30	296	17	0.912	0.948	0.929	0.946	0.908	0.926	0.928

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
4	300	19	307	26	0.940	0.920	0.930	0.922	0.942	0.932	0.931
5	304	25	301	22	0.924	0.933	0.928	0.932	0.923	0.928	0.928
Average					0.932	0.928	0.930	0.928	0.932	0.930	0.930

ตารางผนวกที่ 24 เมทริกซ์ความสับสนของตัวแบบเทคนิคป่าสุ่ม ภายใต้ชุดข้อมูลสายพันธุ์ข้าว (สุ่มตัวอย่างแบบง่าย)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	294	19	307	32	0.939	0.902	0.920	0.906	0.942	0.923	0.922
2	306	29	297	20	0.913	0.939	0.926	0.937	0.911	0.924	0.925
3	307	30	296	19	0.911	0.942	0.926	0.940	0.908	0.924	0.925
4	296	19	307	30	0.940	0.908	0.924	0.911	0.942	0.926	0.925
5	305	26	300	21	0.921	0.936	0.928	0.935	0.920	0.927	0.928
Average					0.925	0.925	0.925	0.926	0.925	0.925	0.925

ตารางผนวกที่ 25 แสดงผลลัพธ์การจำแนกโครงข่ายประสาทเทียม ภายใต้ชุดข้อมูลสายพันธุ์ข้าว (สุ่มตัวอย่างแบบง่าย)

(l, m, h)	Yes			No			Acc
	Pre	Re	F	Pre	Re	F	
(0.1, 0.1, 2)	0.933	0.923	0.928	0.923	0.941	0.929	0.930
(0.1, 0.1, 3)	0.932	0.930	0.931	0.933	0.937	0.930	0.928
(0.1, 0.1, 4)	0.937	0.927	0.932	0.929	0.944	0.933	0.931
(0.1, 0.1, 5)	0.928	0.927	0.927	0.926	0.935	0.928	0.929
(0.1, 0.1, 6)	0.938	0.925	0.931	0.928	0.944	0.932	0.930
(0.2, 0.1, 2)	0.933	0.924	0.928	0.924	0.941	0.929	0.930
(0.2, 0.1, 3)	0.932	0.930	0.931	0.933	0.937	0.931	0.928
(0.2, 0.1, 4)	0.938	0.926	0.932	0.929	0.944	0.933	0.931
(0.2, 0.1, 5)	0.933	0.926	0.929	0.925	0.941	0.930	0.931

(l, m, h)	Yes			No			Acc
	Pre	Re	F	Pre	Re	F	
(0.2, 0.1, 6)	0.938	0.925	0.931	0.928	0.944	0.932	0.930
(0.1, 0.2, 2)	0.934	0.928	0.931	0.930	0.940	0.931	0.930
(0.1, 0.2, 3)	0.927	0.931	0.929	0.934	0.932	0.928	0.926
(0.1, 0.2, 4)	0.941	0.923	0.932	0.926	0.948	0.933	0.931
(0.1, 0.2, 5)	**0.933**	**0.926**	**0.929**	**0.924**	**0.941**	**0.930**	**0.933**
(0.1, 0.2, 6)	0.936	0.926	0.931	0.929	0.942	0.931	0.929
(0.2, 0.2, 2)	0.935	0.929	0.932	0.931	0.941	0.932	0.931
(0.2, 0.2, 3)	0.927	0.933	0.929	0.935	0.931	0.929	0.927
(0.2, 0.2, 4)	0.940	0.925	0.932	0.928	0.946	0.933	0.931
(0.2, 0.2, 5)	0.934	0.925	0.929	0.924	0.942	0.930	0.933
(0.2, 0.2, 6)	0.936	0.926	0.931	0.929	0.942	0.931	0.929

ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (สุ่มตัวอย่างแบบง่าย)

ตารางผนวกที่ 26 เมทริกซ์ความสับสนของตัวแบบการวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์ ภายใต้ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (สุ่มตัวอย่างแบบง่าย)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	211	126	170	85	0.626	0.713	0.667	0.667	0.574	0.617	0.644
2	196	123	173	100	0.614	0.662	0.637	0.634	0.584	0.608	0.623
3	195	129	168	102	0.602	0.657	0.628	0.622	0.566	0.593	0.611
4	199	121	176	98	0.622	0.670	0.645	0.642	0.593	0.616	0.631
5	197	143	154	100	0.579	0.663	0.619	0.606	0.519	0.559	0.591
Average					0.609	0.673	0.639	0.634	0.567	0.599	0.620

ตารางผนวกที่ 27 เมทริกซ์ความสับสนของตัวแบบเทคนิคนาอีฟเบย์ ภายใต้ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (สุ่มตัวอย่างแบบง่าย)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	216	163	133	80	0.570	0.730	0.640	0.624	0.449	0.523	0.590
2	196	133	163	100	0.596	0.662	0.627	0.620	0.551	0.583	0.606
3	209	156	141	88	0.573	0.704	0.631	0.616	0.475	0.536	0.589
4	224	178	119	73	0.557	0.754	0.641	0.620	0.401	0.487	0.577
5	217	188	109	80	0.536	0.731	0.618	0.577	0.367	0.449	0.549
Average					0.566	0.716	0.632	0.611	0.448	0.515	0.582

ตารางผนวกที่ 28 เมทริกซ์ความสับสนของตัวแบบต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 ภายใต้ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (สุ่มตัวอย่างแบบง่าย)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	183	127	169	113	0.590	0.618	0.604	0.599	0.571	0.585	0.595
2	196	116	180	100	0.628	0.662	0.645	0.643	0.608	0.625	0.635
3	184	123	174	113	0.599	0.620	0.609	0.606	0.586	0.596	0.603
4	186	109	188	111	0.631	0.626	0.628	0.629	0.633	0.631	0.630
5	177	124	173	120	0.588	0.596	0.592	0.590	0.582	0.586	0.589
Average					0.607	0.624	0.616	0.614	0.596	0.605	0.610

ตารางผนวกที่ 29 เมทริกซ์ความสับสนของตัวแบบเทคนิคป่าสุ่ม ภายใต้ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (สุ่มตัวอย่างแบบง่าย)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	186	115	181	110	0.618	0.628	0.623	0.622	0.611	0.617	0.620
2	187	118	178	109	0.613	0.632	0.622	0.620	0.601	0.611	0.617
3	184	112	185	113	0.622	0.620	0.621	0.621	0.623	0.622	0.621
4	181	113	184	116	0.616	0.609	0.613	0.613	0.620	0.616	0.614

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
5	194	135	162	103	0.590	0.653	0.620	0.611	0.545	0.577	0.599
Average					0.612	0.628	0.620	0.618	0.600	0.608	0.614

ตารางผนวกที่ 30 แสดงผลลัพธ์การจำแนกโครงข่ายประสาทเทียม ภายใต้ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (สุ่มตัวอย่างแบบง่าย**)**

(l, m, h)	Yes			No			Acc
	Pre	Re	F	Pre	Re	F	
(0.1, 0.1, 13)	0.603	0.666	0.632	0.629	0.561	0.592	0.614
(0.1, 0.1, 14)	0.608	0.666	0.634	0.630	0.566	0.593	0.616
(0.1, 0.1, 15)	0.617	0.628	0.620	0.623	0.608	0.613	0.618
(0.1, 0.1, 16)	0.611	0.639	0.616	0.620	0.578	0.589	0.609
(0.1, 0.1, 17)	0.638	0.530	0.579	0.598	0.699	0.644	0.614
(0.2, 0.1, 13)	0.607	0.662	0.632	0.631	0.571	0.598	0.617
(0.2, 0.1, 14)	0.608	0.667	0.634	0.630	0.566	0.594	0.616
(0.2, 0.1, 15)	0.617	0.625	0.619	0.622	0.611	0.614	0.618
(0.2, 0.1, 16)	0.610	0.646	0.618	0.623	0.571	0.585	0.609
(0.2, 0.1, 17)	0.644	0.510	0.569	0.594	0.716	0.649	0.613
(0.1, 0.2, 13)	0.611	0.661	0.633	0.633	0.578	0.602	0.619
(0.1, 0.2, 14)	0.612	0.672	0.638	0.634	0.568	0.597	0.620
(0.1,0.2,15)	**0.627**	**0.616**	**0.617**	**0.625**	**0.629**	**0.623**	**0.622**
(0.1, 0.2, 16)	0.615	0.626	0.609	0.618	0.589	0.592	0.608
(0.1, 0.2, 17)	0.639	0.506	0.564	0.591	0.713	0.646	0.610
(0.2, 0.2, 13)	0.613	0.657	0.632	0.633	0.585	0.605	0.621
(0.2, 0.2, 14)	0.611	0.667	0.636	0.632	0.571	0.597	0.619
(0.2, 0.2, 15)	0.629	0.606	0.613	0.623	0.638	0.626	0.622
(0.2, 0.2, 16)	0.613	0.631	0.610	0.617	0.581	0.587	0.606
(0.2, 0.2, 17)	0.644	0.499	0.562	0.590	0.722	0.649	0.611

ชุดข้อมูลสถาบันการเงิน (แบ่งกลุ่มข้อมูลแบบเคมีน)

ตารางผนวกที่ 31 เมทริกซ์ความสับสนของตัวแบบการวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์ ภายใต้ชุดข้อมูลสถาบันการเงิน (แบ่งกลุ่มข้อมูลแบบเคมีน)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	651	129	642	120	0.835	0.844	0.839	0.843	0.833	0.838	0.839
2	650	118	654	122	0.846	0.842	0.844	0.843	0.847	0.845	0.845
3	647	127	645	125	0.836	0.838	0.837	0.838	0.835	0.837	0.837
4	651	115	657	121	0.850	0.843	0.847	0.844	0.851	0.848	0.847
5	643	99	673	129	0.867	0.833	0.849	0.839	0.872	0.855	0.852
Average					0.847	0.840	0.843	0.841	0.848	0.844	0.844

ตารางผนวกที่ 32 เมทริกซ์ความสับสนของตัวแบบเทคนิคนาอีฟเบย์ ภายใต้ชุดข้อมูลสถาบันการเงิน (แบ่งกลุ่มข้อมูลแบบเคมีน)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	432	79	692	339	0.845	0.560	0.674	0.671	0.898	0.768	0.729
2	522	156	616	250	0.770	0.676	0.720	0.711	0.798	0.752	0.737
3	539	142	630	233	0.791	0.698	0.742	0.730	0.816	0.771	0.757
4	530	144	628	242	0.786	0.687	0.733	0.722	0.813	0.765	0.750
5	530	142	630	242	0.789	0.687	0.734	0.722	0.816	0.766	0.751
Average					0.796	0.662	0.721	0.711	0.828	0.764	0.745

ตารางผนวกที่ 33 เมทริกซ์ความสับสนของตัวแบบต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 ภายใต้ชุดข้อมูลสถาบันการเงิน (แบ่งกลุ่มข้อมูลแบบเคมีน)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	708	156	615	63	0.819	0.918	0.866	0.907	0.798	0.849	0.858
2	693	139	633	79	0.833	0.898	0.864	0.889	0.820	0.853	0.859
3	690	145	627	82	0.826	0.894	0.859	0.884	0.812	0.847	0.853

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
4	696	144	628	76	0.829	0.902	0.864	0.892	0.813	0.851	0.858
5	685	127	645	87	0.844	0.887	0.865	0.881	0.835	0.858	0.861
Average					0.830	0.900	0.863	0.891	0.816	0.851	0.858

ตารางผนวกที่ 34 เมทริกซ์ความสับสนของตัวแบบเทคนิคป่าสุ่ม ภายใต้ชุดข้อมูลสถาบันการเงิน (แบ่งกลุ่มข้อมูลแบบเคมีน)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	709	137	634	62	0.838	0.920	0.877	0.911	0.822	0.864	0.871
2	708	131	641	64	0.844	0.917	0.879	0.909	0.830	0.868	0.874
3	707	145	627	65	0.830	0.916	0.871	0.906	0.812	0.857	0.864
4	710	136	636	62	0.839	0.920	0.878	0.911	0.824	0.865	0.872
5	702	131	641	70	0.843	0.909	0.875	0.902	0.830	0.864	0.870
Average					0.839	0.916	0.876	0.908	0.824	0.864	0.870

ตารางผนวกที่ 35 แสดงผลลัพธ์การจำแนกโครงข่ายประสาทเทียม ภายใต้ชุดข้อมูลสถาบันการเงิน (แบ่งกลุ่มข้อมูลแบบเคมีน)

(l, m, h)	Yes			No			Acc
	Pre	Re	F	Pre	Re	F	
(0.1, 0.1, 21)	0.815	0.926	0.867	0.915	0.789	0.847	0.857
(0.1, 0.1, 22)	0.835	0.890	0.861	0.883	0.823	0.852	0.857
(0.1, 0.1, 23)	0.816	0.924	0.866	0.913	0.791	0.847	0.857
(0.1, 0.1, 24)	0.858	0.839	0.848	0.844	0.860	0.851	0.850
(0.1, 0.1, 25)	0.836	0.889	0.861	0.882	0.825	0.852	0.857
(0.2, 0.1, 21)	0.812	0.927	0.866	0.916	0.785	0.845	0.856
(0.2,0.1,22)	**0.835**	**0.893**	**0.862**	**0.885**	**0.823**	**0.853**	**0.858**
(0.2, 0.1, 23)	0.814	0.924	0.865	0.914	0.788	0.845	0.856
(0.2, 0.1, 24)	0.854	0.847	0.850	0.850	0.854	0.851	0.851

(l, m, h)	Yes			No			Acc
	Pre	Re	F	Pre	Re	F	
(0.2, 0.1, 25)	0.837	0.889	0.862	0.882	0.826	0.853	0.857
(0.1, 0.2, 21)	0.807	0.927	0.863	0.915	0.778	0.840	0.853
(0.1, 0.2, 22)	0.830	0.898	0.862	0.889	0.816	0.850	0.857
(0.1, 0.2, 23)	0.809	0.929	0.864	0.918	0.778	0.842	0.854
(0.1, 0.2, 24)	0.858	0.836	0.846	0.841	0.860	0.850	0.848
(0.1, 0.2, 25)	0.831	0.900	0.864	0.893	0.816	0.852	0.858
(0.2, 0.2, 21)	0.806	0.926	0.862	0.914	0.777	0.839	0.852
(0.2, 0.2, 22)	0.828	0.899	0.862	0.890	0.813	0.849	0.856
(0.2, 0.2, 23)	0.808	0.931	0.865	0.920	0.778	0.842	0.854
(0.2, 0.2, 24)	0.852	0.842	0.847	0.846	0.853	0.849	0.848
(0.2, 0.2, 25)	0.832	0.897	0.862	0.889	0.817	0.851	0.857

ชุดข้อมูลสายพันธุ์ข้าว (แบ่งกลุ่มข้อมูลแบบเคมีน)

ตารางผนวกที่ 36 เมทริกซ์ความสับสนของตัวแบบการวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์ ภายใต้ชุดข้อมูลสายพันธุ์ข้าว (แบ่งกลุ่มข้อมูลแบบเคมีน)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	305	30	296	21	0.910	0.936	0.923	0.934	0.908	0.921	0.922
2	303	23	303	23	0.929	0.929	0.929	0.929	0.929	0.929	0.929
3	309	22	304	17	0.934	0.948	0.941	0.947	0.933	0.940	0.940
4	304	23	303	22	0.930	0.933	0.931	0.932	0.929	0.931	0.931
5	304	24	302	22	0.927	0.933	0.930	0.932	0.926	0.929	0.929
Average					0.926	0.936	0.931	0.935	0.925	0.930	0.930

ตารางผนวกที่ 37 เมทริกซ์ความสับสนของตัวแบบเทคนิคนาอีฟเบย์ ภายใต้ชุดข้อมูลสายพันธุ์ข้าว (แบ่งกลุ่มข้อมูลแบบเคมีน)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	308	34	292	18	0.901	0.945	0.922	0.942	0.896	0.918	0.920
2	305	32	294	21	0.905	0.936	0.920	0.933	0.902	0.917	0.919
3	306	34	292	20	0.900	0.939	0.919	0.936	0.896	0.915	0.917
4	301	34	292	25	0.899	0.923	0.911	0.921	0.896	0.908	0.910
5	300	31	295	26	0.906	0.920	0.913	0.919	0.905	0.912	0.913
Average					0.902	0.933	0.917	0.930	0.899	0.914	0.916

ตารางผนวกที่ 38 เมทริกซ์ความสับสนของตัวแบบต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 ภายใต้ชุดข้อมูลสายพันธุ์ข้าว (แบ่งกลุ่มข้อมูลแบบเคมีน)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	303	33	293	23	0.902	0.929	0.915	0.927	0.899	0.913	0.914
2	301	22	304	25	0.932	0.923	0.928	0.924	0.933	0.928	0.928
3	309	23	303	17	0.931	0.948	0.939	0.947	0.929	0.938	0.939
4	302	21	305	24	0.935	0.926	0.931	0.927	0.936	0.931	0.931
5	300	22	304	26	0.932	0.920	0.926	0.921	0.933	0.927	0.926
Average					0.926	0.929	0.928	0.929	0.926	0.927	0.928

ตารางผนวกที่ 39 เมทริกซ์ความสับสนของตัวแบบเทคนิคป่าสุ่ม ภายใต้ชุดข้อมูลสายพันธุ์ข้าว (แบ่งกลุ่มข้อมูลแบบเคมีน)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	302	33	293	24	0.901	0.926	0.914	0.924	0.899	0.911	0.913
2	307	25	301	19	0.925	0.942	0.933	0.941	0.923	0.932	0.933
3	306	33	293	20	0.903	0.939	0.920	0.936	0.899	0.917	0.919
4	303	25	301	23	0.924	0.929	0.927	0.929	0.923	0.926	0.926

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
5	295	24	302	31	0.925	0.905	0.915	0.907	0.926	0.917	0.916
Average					0.915	0.928	0.922	0.927	0.914	0.921	0.921

ตารางผนวกที่ 40 แสดงผลลัพธ์การจำแนกโครงข่ายประสาทเทียม ภายใต้ชุดข้อมูลสายพันธุ์ข้าว (แบ่งกลุ่มข้อมูลแบบเคมีน)

(l, m, h)	Yes			No			Acc
	Pre	Re	F	Pre	Re	F	
(0.1, 0.1, 2)	0.932	0.919	0.926	0.920	0.933	0.927	0.926
(0.1, 0.1, 3)	0.934	0.920	0.927	0.921	0.935	0.928	0.927
(0.1, 0.1, 4)	0.935	0.917	0.926	0.918	0.936	0.927	0.926
(0.1, 0.1, 5)	0.929	0.928	0.928	0.928	0.929	0.928	0.928
(0.1, 0.1, 6)	0.931	0.921	0.926	0.922	0.931	0.927	0.926
(0.2, 0.1, 2)	0.932	0.920	0.926	0.921	0.933	0.927	0.926
(0.2, 0.1, 3)	0.930	0.920	0.925	0.922	0.931	0.926	0.925
(0.2, 0.1, 4)	0.936	0.917	0.926	0.918	0.937	0.928	0.927
(0.2, 0.1, 5)	0.928	0.928	0.928	0.928	0.928	0.928	0.928
(0.2, 0.1, 6)	0.931	0.921	0.926	0.922	0.931	0.927	0.926
(0.1, 0.2, 2)	0.932	0.919	0.926	0.920	0.933	0.927	0.926
(0.1, 0.2, 3)	0.936	0.914	0.925	0.916	0.937	0.927	0.926
(0.1, 0.2, 4)	0.937	0.912	0.924	0.915	0.939	0.926	0.925
(0.1, 0.2, 5)	0.928	0.927	0.927	0.928	0.928	0.927	0.927
(0.1, 0.2, 6)	**0.933**	**0.921**	**0.927**	**0.923**	**0.934**	**0.928**	**0.928**
(0.2, 0.2, 2)	0.932	0.919	0.925	0.920	0.933	0.926	0.926
(0.2, 0.2, 3)	0.936	0.915	0.925	0.917	0.937	0.927	0.926
(0.2, 0.2, 4)	0.937	0.911	0.924	0.914	0.939	0.926	0.925
(0.2, 0.2, 5)	0.928	0.927	0.927	0.928	0.928	0.927	0.927
(0.2, 0.2, 6)	0.934	0.920	0.927	0.922	0.935	0.928	0.928

ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (แบ่งกลุ่มข้อมูลแบบเคมีน)

ตารางผนวกที่ 41 เมทริกซ์ความสับสนของตัวแบบการวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์ ภายใต้ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (แบ่งกลุ่มข้อมูลแบบเคมีน)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	203	137	159	93	0.597	0.686	0.638	0.631	0.537	0.580	0.611
2	207	137	159	89	0.602	0.699	0.647	0.641	0.537	0.585	0.618
3	202	119	178	95	0.629	0.680	0.654	0.652	0.599	0.625	0.640
4	187	127	170	110	0.596	0.630	0.612	0.607	0.572	0.589	0.601
5	206	122	175	91	0.628	0.694	0.659	0.658	0.589	0.622	0.641
Average					0.610	0.678	0.642	0.638	0.567	0.600	0.622

ตารางผนวกที่ 42 เมทริกซ์ความสับสนของตัวแบบเทคนิคนาอีฟเบย์ ภายใต้ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (แบ่งกลุ่มข้อมูลแบบเคมีน)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	211	169	127	85	0.555	0.713	0.624	0.599	0.429	0.500	0.571
2	215	153	143	81	0.584	0.726	0.648	0.638	0.483	0.550	0.605
3	191	145	152	106	0.568	0.643	0.603	0.589	0.512	0.548	0.577
4	207	175	122	90	0.542	0.697	0.610	0.575	0.411	0.479	0.554
5	215	169	128	82	0.560	0.724	0.631	0.610	0.431	0.505	0.577
Average					0.562	0.701	0.623	0.602	0.453	0.516	0.577

ตารางผนวกที่ 43 เมทริกซ์ความสับสนของตัวแบบต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 ภายใต้ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (แบ่งกลุ่มข้อมูลแบบเคมีน)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	164	128	168	132	0.562	0.554	0.558	0.560	0.568	0.564	0.561
2	186	121	175	110	0.606	0.628	0.617	0.614	0.591	0.602	0.610
3	192	134	163	105	0.589	0.646	0.616	0.608	0.549	0.577	0.598

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
4	176	120	177	121	0.595	0.593	0.594	0.594	0.596	0.595	0.594
5	171	117	180	126	0.594	0.576	0.585	0.588	0.606	0.597	0.591
Average					0.589	0.599	0.594	0.593	0.582	0.587	0.591

ตารางผนวกที่ 44 เมทริกซ์ความสับสนของตัวแบบเทคนิคป่าสุ่ม ภายใต้ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (แบ่งกลุ่มข้อมูลแบบเคมีน)

Fold	TP	FP	TN	FN	Yes			No			Acc
					Pre	Re	F	Pre	Re	F	
1	186	130	166	110	0.589	0.628	0.608	0.601	0.561	0.580	0.595
2	197	125	171	99	0.612	0.666	0.638	0.633	0.578	0.604	0.622
3	194	119	178	103	0.620	0.653	0.636	0.633	0.599	0.616	0.626
4	178	120	177	119	0.597	0.599	0.598	0.598	0.596	0.597	0.598
5	180	118	179	117	0.604	0.606	0.605	0.605	0.603	0.604	0.604
Average					0.604	0.631	0.617	0.614	0.587	0.600	0.609

ตารางผนวกที่ 45 แสดงผลลัพธ์การจำแนกโครงข่ายประสาทเทียม ภายใต้ชุดข้อมูลนักวิทยาศาสตร์ข้อมูล (แบ่งกลุ่มข้อมูลแบบเคมีน)

(l, m, h)	Yes			No			Acc
	Pre	Re	F	Pre	Re	F	
(0.1, 0.1, 13)	0.608	0.695	0.649	0.644	0.552	0.594	0.623
(0.1, 0.1, 14)	0.603	0.691	0.644	0.639	0.545	0.588	0.618
(0.1, 0.1, 15)	0.615	0.648	0.629	0.629	0.591	0.607	0.620
(0.1, 0.1, 16)	0.606	0.651	0.624	0.628	0.575	0.596	0.613
(0.1, 0.1, 17)	0.632	0.564	0.595	0.606	0.670	0.635	0.617
(0.2, 0.1, 13)	0.607	0.695	0.647	0.642	0.549	0.592	0.622
(0.2, 0.1, 14)	0.603	0.695	0.646	0.641	0.542	0.587	0.619
(0.2, 0.1, 15)	0.616	0.647	0.629	0.628	0.593	0.607	0.620
(0.2, 0.1, 16)	0.606	0.664	0.630	0.632	0.565	0.592	0.615

(l, m, h)	Yes			No			Acc
	Pre	Re	F	Pre	Re	F	
(0.2, 0.1, 17)	0.631	0.562	0.594	0.602	0.668	0.633	0.615
(0.1,0.2,13)	**0.611**	**0.692**	**0.649**	**0.645**	**0.559**	**0.598**	**0.625**
(0.1, 0.2, 14)	0.600	0.698	0.645	0.640	0.535	0.582	0.616
(0.1, 0.2, 15)	0.613	0.655	0.630	0.630	0.582	0.601	0.618
(0.1, 0.2, 16)	0.609	0.655	0.627	0.631	0.576	0.597	0.615
(0.1, 0.2, 17)	0.629	0.558	0.590	0.603	0.670	0.634	0.614
(0.2,0.2,13)	0.611	0.690	0.648	0.645	0.561	0.599	0.625
(0.2, 0.2, 14)	0.600	0.698	0.645	0.640	0.534	0.582	0.616
(0.2, 0.2, 15)	0.614	0.651	0.629	0.629	0.584	0.602	0.618
(0.2, 0.2, 16)	0.605	0.660	0.627	0.630	0.565	0.590	0.612
(0.2, 0.2, 17)	0.624	0.552	0.585	0.599	0.667	0.631	0.610

ภาคผนวก ข

โปรแกรมอาร์

การสร้างตัวแบบจำแนกด้วยการวิเคราะห์จำแนกกลุ่มเชิงเส้นโดยวิธีของฟิชเชอร์

```
library(MASS)
library('caret')
for (i in 1:5) {
  set.seed(8830)
  test = subset(data, k == i)
  train = subset(data, k != i)
  test = test[,2:31]
  train = train[,2:31]
  fit.lda = lda(target  ~ ., data = train)
  pred.lda = predict(fit.lda, newdata = test[,2:30])
  print(confusionMatrix(pred.lda$class, test$target , positive = 'yes'))
}
```

การสร้างตัวแบบจำแนกด้วยเทคนิคนาอีฟเบย์

```
library(naivebayes)
library('caret')
for (i in 1:5) {
  test = subset(data, k == i)
  train = subset(data, k != i)
  test = test[,2:31]
  train = train[,2:31]
  fit.ann = naive_bayes(target ~., data=train)
  pred <- predict(fit.ann,newdata = test[,2:30])
  print(confusionMatrix(pred, test$target, positive = 'yes'))
}
```

การสร้างตัวแบบจำแนกด้วยต้นไม่ตัดสินใจด้วยอัลกอริทึม C4.5

```
library(RWeka)
library('caret')
for (i in 1:5) {
  set.seed(10)
  test = subset(data, k == i)
```

```
  train = subset(data, k != i)
  test = test[,2:31]
  train = train[,2:31]
  fit.J48 = J48(target ~. , data = train)
  pred = predict(fit.J48, newdata = test[,2:30])
  print(confusionMatrix(pred, test$target, positive = 'yes'))
  }
```

การสร้างตัวแบบจำแนกด้วยเทคนิคป่าสุ่ม

```
library(randomForest)
library('caret')
for (i in 1:5) {
  set.seed(10)
  print(i)
  test = subset(data, k == i)
  train = subset(data, k != i)
  test = test[,2:31]
  train = train[,2:31]
  fit.RF = randomForest(target~., data = train, importance=TRUE,proximity=TRUE)
  pred <- predict(fit.RF,newdata = test[,2:30])
  print(confusionMatrix(pred, test$target, positive = 'yes'))
  }
```

การสร้างตัวแบบจำแนกด้วยโครงข่ายประสาทเทียม

```
library(ANN2)
library('caret')
k_fold = c(1, 2, 3, 4, 5)
learning_rate = c(0.1, 0.2)
momentum = c(0.1, 0.2)
hidden_layer = c(13, 14, 15, 16, 17)

for (i in k_fold) {
  print(paste(i, "is k-fold"))
```

```
for (j in learning_rate) {
  for (k in momentum) {
    for (l in hidden_layer) {
      print(paste("(", k, j, l, ")"))
      #print(paste(j, "is learning_rate"))
      #print(paste(k, "is momentum"))
      #print(paste(l, "is hidden_layer"))
      test = subset(data, k == i)
      train = subset(data, k != i)
      test = test[,2:31]
      train = train[,2:31]
      X_test = test[,2:30]
      y_test = test[,1]
      X_train = train[,2:30]
      y_train = train[,1]
      fit.ann = neuralnetwork(X = X_train,
                              y = y_train,
                              sgd.momentum = k,
                              learn.rates = j,
                              hidden.layers = l,
                              activ.functions = "sigmoid",
                              standardize = FALSE,
                              random.seed = 10,
                              n.epochs = 100)
      pred <- predict(fit.ann ,newdata = X_test)
      print(table(pred$predictions, test$target))
      print(confusionMatrix(pred$predictions, test$target, positive = 'yes'))
    }}}}
```

บรรณานุกรม

Aziz Mohammad Nasrul, & Ahmad Tohari. (2021). Clustering under-sampling data for improving the performance of intrusion detection system. *Journal of Engineering Science and Technology*, *16*(2), 1342-1355.

Burk, S., & Miner, G. D. (2020). *It's All Analytics!: The Foundations of AI, Big Data, and Data Science Landscape for Professionals in Healthcare, Business, and Government*. Productivity Press.

Cinar, I., & Koklu, M. (2019). Classification of rice varieties using artificial intelligence methods. *International Journal of Intelligent Systems and Applications in Engineering*, *7*(3), 188-194.

Comert, Z., & Kocamaz, A. (2017). Comparison of machine learning techniques for fetal heart rate classification.

Cui, M. (2020). Introduction to the k-means clustering algorithm based on the elbow method. *Accounting, Auditing and Finance*, *1*(1), 5-8.

Golpour, P., Ghayour-Mobarhan, M., Saki, A., Esmaily, H., Taghipour, A., Tajfard, M., Ghazizadeh, H., Moohebati, M., & Ferns, G. A. (2020). Comparison of support vector machine, naïve Bayes and logistic regression for assessing the necessity for coronary angiography. *International journal of environmental research and public health*, *17*(18), 6449.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and techniques* (3 ed.). CA: Morgan Kaufmann.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, *21*(9), 1263-1284.

Hutapea, J. Y., Samuel, Y. T., & Sitorus, H. (2019). Comparison of Accuracy Between Two Methods: Naive Bayes Algorithm and Decision Tree-J48 to Predict The Stock Price of Pt Astra International tbk Using Data From Indonesia Stock Exchange. Abstract Proceedings International Scholars Conference,

Johnson, R. (2014). *Applied multivariate statistical analysis* (6th ed ed.). Pearson.

Johnson, R. A., & Wichern, D. W. (2014). *Applied multivariate statistical analysis* (Vol. 6).

Pearson London, UK:.

Kublanov, V. S., Dolganov, A. Y., Belo, D., & Gamboa, H. (2017). Comparison of machine learning methods for the arterial hypertension diagnostics. *Applied bionics and biomechanics*, *2017*.

Pakdaman, M., Naghab, S. S., Khazanedari, L., Malbousi, S., & Falamarzi, Y. (2020). Lightning prediction using an ensemble learning approach for northeast of Iran. *Journal of Atmospheric and Solar-Terrestrial Physics*, *209*, 105417.

Religia, Y., Pranoto, G. T., & Santosa, E. D. (2020). South German Credit Data Classification Using Random Forest Algorithm to Predict Bank Credit Receipts. *JISA (Jurnal Informatika dan Sains)*, *3*(2), 62-66.

Robin, G., & Jean-Michel, P. (2020). *Random Forests with R* [Book]. Springer. https://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=edsebk&AN=2618467&site=eds-live&custid=ns004377

Tarakci, F., & Ozkan, I. A. (2021). Comparison of classification performance of kNN and WKNN algorithms. *Selcuk University Journal of Engineering Sciences*, *20*(2), 32-37.

Thanh Noi, P., & Kappas, M. (2017). Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery. *Sensors*, *18*(1), 18.

Verma, A. (2019). Evaluation of classification algorithms with solutions to class imbalance problem on bank marketing dataset using WEKA. *International Research Journal of Engineering and Technology*, *5*(13), 54-60.

กาญจน์ ณ ศรีธะ. (2560). การเปรียบเทียบเทคนิคการสุ่มตัวอย่างเพื่อการจำแนกข้อมูลที่ไม่สมดุล. *Journal of Applied Informatics and Technology*, *1*(1), 20-37.

พัชรียา ทองพูล. (2562). การเปรียบเทียบประสิทธิภาพในการทำนายผลการปรับความไม่สมดุลของข้อมูลในการจำแนกด้วยเทคนิคการทำเหมืองข้อมูล. *Thai Journal of Science and Technology*, *8*(6), 565-584.

พัฒนพงษ์ ดลรัตน์. (2560). การเปรียบเทียบประสิทธิภาพของแบบจำลองในการพยากรณ์ความสำเร็จการศึกษาของนักเรียนระดับประกาศนียบัตรวิชาชีพ. สารวิทยาศาสตร์และเทคโนโลยี, *37*, 380-338.

วัชรีวรรณ จิตต์สกุล. (2560). การวิเคราะห์การจำแนกข้อความด้วยการเปรียบเทียบความเสถียรของ

อัลกอริทึม. *Sripatum Review of Science and Technology*, *9*(1), 19-31.

วิชญ์วิสฐ เกษรสิทธิ์. (2561). การแก้ไขปัญหาข้อมูลไม่สมดุลของข้อมูลสำหรับการจำแนกผู้ป่วยโรคเบาหวาน. *18*.

สห ธิติถามวัต. (2562). การพัฒนารูปแบบการแนะนำงานสำหรับองค์กรและผู้สมัครตามทักษะการเรียนรู้ด้วยเทคนิคป่าแบบสุ่ม สถาบันเทคโนโลยีไทย-ญี่ปุ่น].

สายชล สินสมบูรณ์ทอง. (2560). การทำเหมืองข้อมูลเล่ม *1* การค้นหาความรู้จากข้อมูล (พิมพ์ครั้งที่ 2 ed.). จามจุรีโปรดักส์.

สุกฤษฎิ์ ไกรนรา. (2563). การศึกษาจำนวนและตำแหน่งสถานีขนส่งสินค้าทางรางที่เหมาะสม กรณีศึกษาการขนส่งน้ำตาลในภาคตะวันออกเฉียงเหนือ มหาวิทยาลัยเกษตรศาสตร์].

อกนิษฐ์ ทองจิตร. (2562). การพัฒนาวิธีจำแนกประเภทข้อมูลโดยใช้โครงข่ายประสาทเทียมแบบปรับเหมาะผสมผสานการหาค่าเหมาะสมที่สุดแบบกลุ่มอนุภาค สำหรับการจำแนกประเภทกลุ่มเสี่ยง. วิทยาการวิจัยและวิทยาการปัญญา, *17*.