การจำแนกข้อมูลขนาดใหญ่มากโดยใช้การจัดกลุ่มด้วยวิธีเคมีนและวิธีการเรียนรู้เชิงลึก

วิทยานิพนธ์

ของ

นันทัชพร เสนาวงค์

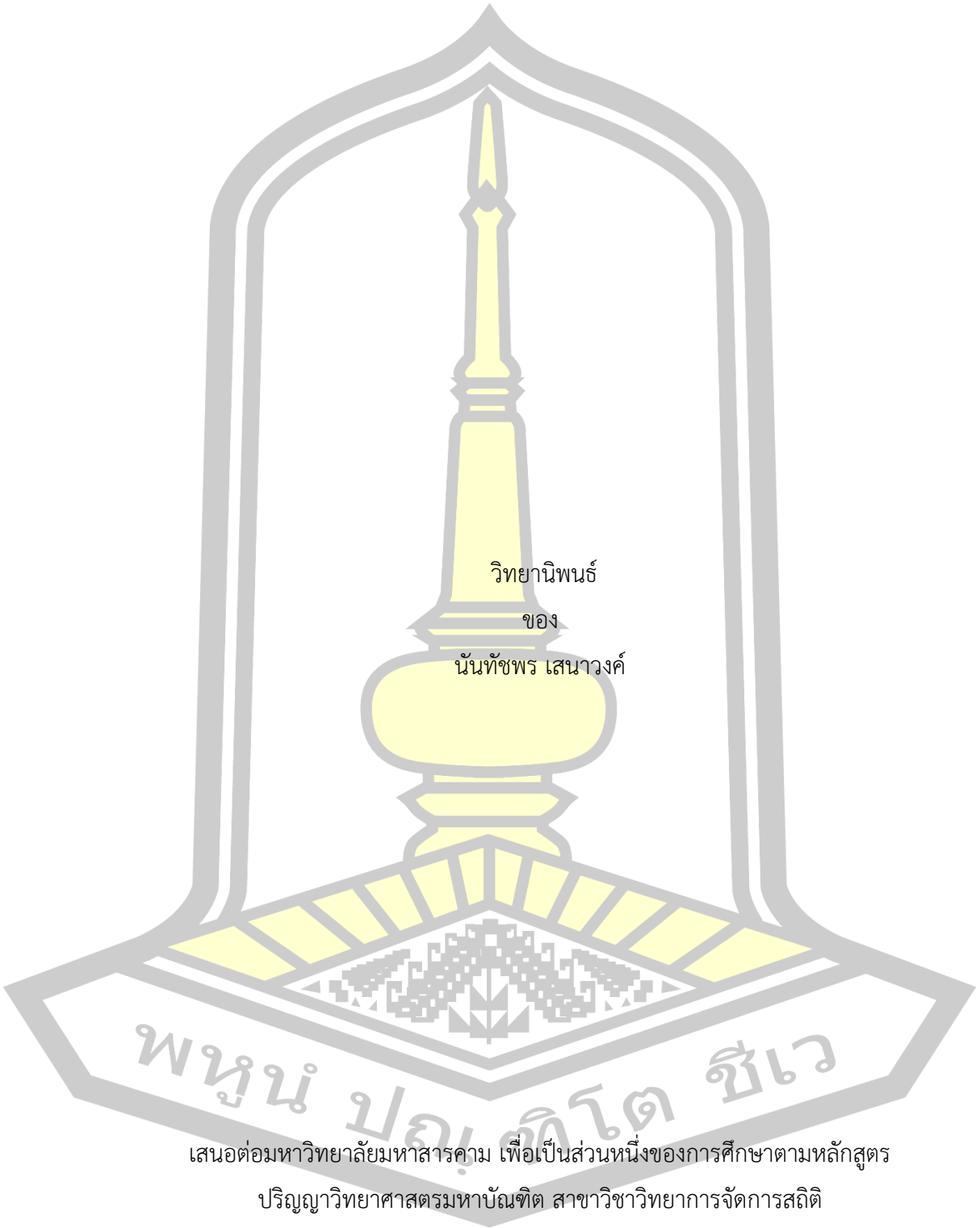
เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการจัดการสถิติ

ธันวาคม 2563

พหุน ปณฺ ฑิโต ชีเว

การจำแนกข้อมูลขนาดใหญ่มากโดยใช้การจัดกลุ่มด้วยวิธีเคมีนและวิธีการเรียนรู้เชิงลึก



วิทยานิพนธ์

ของ

นันทัชพร เสนาวงค์

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการจัดการสถิติ

ธันวาคม 2563

Very Large-scale Data Classification based on K-means Clustering and Deep Learning

Nuntuschaporn Senawong

A Thesis Submitted in Partial Fulfillment of Requirements

for Master of Science (Statistical Management Science)

December 2020

คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาวิทยานิพนธ์ของนางสาวนันทัชพร เสนาวงค์ แล้วเห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการจัดการสถิติ ของมหาวิทยาลัยมหาสารคาม

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(อ. ดร. อาทิตย์ อภิโชติธนกุล)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(รศ. ดร. อรวิชญ์ กุมพล)

.....กรรมการ
(อ. ดร. โรจนี หอมชาลี)

.....กรรมการ
(ผศ. ดร. มนชยา เจียงประดิษฐ์)

มหาวิทยาลัยอนุมัติให้รับวิทยานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการจัดการสถิติ ของมหาวิทยาลัยมหาสารคาม

.....
(ศ. ดร. ไพโรจน์ ประมวล)
คณบดีคณะวิทยาศาสตร์

.....
(รศ. ดร. กริสน์ ชัยมูล)
คณบดีบัณฑิตวิทยาลัย

ชื่อเรื่อง การจำแนกข้อมูลขนาดใหญ่มากโดยใช้การจัดกลุ่มด้วยวิธีเคมีนและวิธีการเรียนรู้เชิงลึก

ผู้วิจัย นันทัชพร เสนาวงค์

อาจารย์ที่ปรึกษา รองศาสตราจารย์ ดร. อรวิชญ์ กุมพล

ปริญญา วิทยาศาสตรมหาบัณฑิต **สาขาวิชา** วิทยาการจัดการสถิติ

มหาวิทยาลัย มหาวิทยาลัยมหาสารคาม **ปีที่พิมพ์** 2563

บทคัดย่อ

ในการจำแนกประเภทข้อมูลที่มีขนาดใหญ่มาก ปัญหาที่พบคือเวลาที่ใช้ในการประมวลผลนาน และต้องใช้ข้อมูลฝึก (Training Data) เป็นจำนวนมากเพื่อให้การจำแนกประเภทมีประสิทธิภาพความแม่นยำสูง เพื่อแก้ไขปัญหานี้ผู้วิจัยจึงศึกษาวิธีการสำหรับการจำแนกข้อมูลขนาดใหญ่มาก เพื่อลดปัญหาการใช้ข้อมูลฝึกจำนวนมาก แต่ยังคงมีประสิทธิภาพในการจำแนกประเภทสูง โดยจะทำการลดขนาดข้อมูลฝึกด้วยการรวมเทคนิคการจัดกลุ่มของวิธีเคมีน (K-means) และวิธีการเรียนรู้เชิงลึก (Deep Learning) ในการศึกษาประสิทธิภาพของวิธีการที่นำเสนอพิจารณาจากค่าความแม่นยำและค่า AUC นอกจากนี้ได้ทำการเปรียบเทียบกับวิธีการเรียนรู้เชิงลึกแบบเดิมที่ใช้ข้อมูลฝึกขนาด 80% - 90% ของข้อมูลทั้งหมด และกรณีที่ใช้ข้อมูลฝึกจำนวนเท่ากัน ผลการศึกษาพบว่าวิธีการที่นำเสนอสามารถลดขนาดของข้อมูลฝึกได้อย่างมาก (น้อยกว่า 1% ของขนาดข้อมูลทั้งหมด) แต่ยังคงมีประสิทธิภาพในการจำแนกประเภทสูง และเวลาที่ใช้ในการจำแนกประเภทน้อยกว่าวิธีการเรียนรู้เชิงลึกอย่างมาก

คำสำคัญ : ข้อมูลขนาดใหญ่มาก, การจัดกลุ่มของวิธีเคมีน, การตรวจหาค่าผิดปกติ, วิธีการเรียนรู้เชิงลึก, การจำแนกประเภท

TITLE Very Large-scale Data Classification based on K-means Clustering and Deep Learning

AUTHOR Nuntuschaporn Senawong

ADVISORS Associate Professor Orawich Kumphon , Ph.D.

DEGREE Master of Science **MAJOR** Statistical Management Science

UNIVERSITY Mahasarakham University **YEAR** 2020

ABSTRACT

In classifying very large data, problems are long processing time and it requires a lot of training data in order to maintain high accuracy. To solve these problems, researchers study methods for classifying very large data to reduce the use of large amounts of training data but still have high classification efficiency. The proposed method reduces the size of the training data by combining K-means and deep learning. To study the effectiveness of the proposed method, the accuracy and AUC values were determined. In addition, it was compared with the original deep learning method using training data about 80% - 90% of data and compared with original deep learning using the same amount of training data with the proposed method. The results show that the proposed method is able to significantly reduce the size of the training data (less than 1% of the total data size) but still, have highly effective in classification and the time it takes to classify is significantly less than the deep learning method.

Keyword : Very Large-scale Data, K-means Clustering, Outlier Detection, Deep Learning, Classification

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงด้วยดี ด้วยความกรุณาในการให้คำปรึกษาแนะนำเป็นอย่างดียิ่ง จากรองศาสตราจารย์ ดร.อรวิชญ์ กุมพล อาจารย์ที่ปรึกษาวิทยานิพนธ์ และ อาจารย์ ดร.สุภาวดี วิชิตชาญ ที่ได้ชี้แนวทางในการศึกษาวิทยานิพนธ์ ให้คำแนะนำในการค้นคว้าวิเคราะห์ข้อมูล ตลอดจนแก้ไขปรับปรุงข้อบกพร่องต่าง ๆ ทำให้วิทยานิพนธ์สำเร็จสมบูรณ์ด้วยดี รวมทั้ง อาจารย์ ดร.อาทิตย์ อภิโชติธนกุล ประธานกรรมการสอบวิทยานิพนธ์ อาจารย์ ดร.โรจนี หอมชาลี และ ผู้ช่วยศาสตราจารย์ ดร.มนชยา เจียงประดิษฐ์ กรรมการสอบวิทยานิพนธ์ที่กรุณาให้คำแนะนำและให้ข้อเสนอแนะ ตลอดจนแก้ไขปรับปรุงข้อบกพร่องต่าง ๆ อันเป็นประโยชน์ต่อวิทยานิพนธ์ ทำให้วิทยานิพนธ์ฉบับนี้สมบูรณ์ยิ่งขึ้น ผู้วิจัยขอกราบขอบพระคุณอย่างสูงยิ่งไว้ ณ โอกาสนี้ด้วย

ทั้งนี้ผู้วิจัยขอขอบคุณฐานเก็บข้อมูล KEEL และ UCL ที่ได้อนุเคราะห์ข้อมูลขนาดใหญ่มาก ซึ่งเป็นประโยชน์อย่างมากต่อการทำวิทยานิพนธ์ในครั้งนี้

ผู้วิจัยขอกราบขอบพระคุณบุพการี ซึ่งเป็นผู้ให้ชีวิต ความรัก ความอบอุ่น และบูรพคณาจารย์ ซึ่งเป็นผู้ให้ความรู้ มีส่วนวางรากฐานทางการศึกษา รวมถึงพี่น้องสาขาวิทยาการจัดการสถิติ และผู้มีส่วนเกี่ยวข้องทุกท่าน ซึ่งเป็นผู้ให้แรงบันดาลใจในการทำงาน คอยเตือนสติ และเป็นกำลังใจให้แก่ผู้ทำวิทยานิพนธ์จนบรรลุผลสำเร็จไปได้ด้วยดี

นันทัชพร เสนาวงค์

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1 หลักการและเหตุผล

ในปัจจุบันนิยมใช้วิธีการเรียนรู้ของเครื่อง (Machine Learning) เพื่อสืบค้นความรู้จากฐานข้อมูลขนาดใหญ่ ซึ่งสามารถสกัดข้อมูลที่มีประโยชน์และน่าสนใจ อีกทั้งยังจดจำรูปแบบจากฐานข้อมูลขนาดใหญ่ได้ (Knowledge Discovery from Very Large Databases : KDD) หรือเรียกว่า การทำเหมืองข้อมูล (Data Mining) ซึ่งเป็นเทคนิคที่ใช้จัดการกับข้อมูลขนาดใหญ่ โดยจะนำข้อมูลที่มีมาทำการวิเคราะห์แล้วดึงความรู้หรือสิ่งสำคัญออกมาเพื่อใช้ในการวิเคราะห์หรือทำนายสิ่งต่าง ๆ ที่จะเกิดขึ้น โดยการค้นหาความรู้และความจริงที่แฝงอยู่ในข้อมูล (Knowledge Discovery) เป็นกระบวนการขุดค้นสิ่งที่น่าสนใจในข้อมูลที่มีอยู่ [1] ดังนั้นวิธีการเรียนรู้ของเครื่องจึงเป็นหัวใจหลักในการวิเคราะห์ข้อมูล เช่น การจำแนกประเภทข้อมูลขนาดใหญ่ โดยการใช้เทคนิคการจำแนกประเภทข้อมูล (Data Classification) เพื่อจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนด จากกลุ่มตัวอย่างข้อมูลที่เรียกว่าข้อมูลฝึก (Training Data) ทั้งนี้เมื่อต้องการจำแนกประเภทข้อมูลที่มีขนาดใหญ่ปัญหาที่ตามมาคือการประมวลผลซึ่งต้องใช้เวลานาน และต้องใช้ข้อมูลฝึกเป็นจำนวนมาก โดยทั่วไปจะใช้ข้อมูลฝึกประมาณ 80% ถึง 90% ของข้อมูลทั้งหมด เพื่อได้ประสิทธิภาพความแม่นที่สูง [2] ในการแก้ปัญหาดังกล่าว [3] ได้เสนอวิธีการสำหรับการจำแนกข้อมูลขนาดใหญ่มากโดยใช้การวิเคราะห์กลุ่มด้วยวิธีเคมีน (K-means) และวิธีมัลติเคอร์เนลซัพพอร์ตเวกเตอร์แมชชีน (Multi-kernel Support Vector Machine : Multi-kernel SVM) เพื่อลดขนาดของข้อมูลฝึกและลดเวลาในการประมวลผล ผลลัพธ์ที่ได้แสดงให้เห็นว่าวิธีการเลือกข้อมูลฝึกที่นำเสนอสามารถลดขนาดของข้อมูลฝึกอีกทั้งยังลดเวลาที่ใช้ในการประมวลผลและสามารถรักษาประสิทธิภาพความแม่นได้ดี

ปัจจุบันใช้การจำแนกประเภทด้วยวิธีการเรียนรู้ของเครื่องในข้อมูลที่มีขนาดใหญ่มีหลายวิธี เช่น วิธีการเรียนรู้โดยการใช้โครงข่ายประสาทเทียม (Artificial Neural Network : ANN) โดย [4] ได้ศึกษาการประยุกต์ ANN หลายโครงข่ายบนข้อมูลขนาดใหญ่ แต่มักเกิดปัญหาในด้านเวลาที่ใช้ในการเรียนรู้ จึงได้เสนอการพัฒนาอัลกอริทึมในการรวมโหนด ผลการทดลองพบว่าวิธีการที่นำเสนอสามารถลดเวลาในการเรียนรู้ลงได้อย่างมาก และยังคงรักษาประสิทธิภาพความแม่นได้เหมือนกับการใช้ข้อมูลฝึกขนาด 80% ถึง 90% ทั้งนี้วิธีการเรียนรู้โดยการใช้ ANN ได้พัฒนาเป็น

วิธีการเรียนรู้เชิงลึก (Deep Learning) โดย [5] ได้เสนอการใช้เทคนิคการเรียนรู้เชิงลึกผ่านชุดโครงข่ายประสาทเทียม โดยศึกษาแบบจำลองทำนายผลคำตัดสินและประเด็นในคดีอาญาที่เรียนรู้จากคำพิพากษาศาลฎีกาไทย ผลการทดลองแสดงให้เห็นว่าแบบจำลองที่เสนอให้ประสิทธิภาพความแม่นสูงกว่าแบบจำลองที่ใช้วิธีการเรียนรู้ของเครื่องแบบเดิม เช่น Naive Bayes และ SVM

วิธีการเรียนรู้เชิงลึกพัฒนาและรองรับการวิเคราะห์ข้อมูลที่มีรูปแบบหลากหลายและไร้โครงสร้างได้ดีกว่าสถิติวิเคราะห์แบบเดิม ซึ่งมีความยืดหยุ่นมากกว่าแบบจำลองทางสถิติที่มักจะเป็นแบบจำลองเชิงเส้นตรง โดยที่วิธีการเรียนรู้เชิงลึกเป็นเส้นโค้งหรือเส้นรูปแบบอื่น ๆ และมีความซับซ้อนมากกว่า [6] และยังเป็นเครื่องมือที่ดีที่สุดในปัจจุบันเพื่อวิเคราะห์หารูปแบบของข้อมูล อีกทั้งคอมพิวเตอร์ยังสามารถถูกฝึกได้อย่างอัตโนมัติ วัตถุประสงค์หลักเป็นการใช้ข้อมูลฝึกเพื่อจำแนกประเภทของชนิดในวัตถุนั้น เช่น การทำนายประเภทของวัตถุในวิธีการเรียนรู้ของเครื่องจะทำนายประเภทของวัตถุนั้นว่าอยู่ประเภทใด [7] เมื่อเครื่องมือสามารถทำนายผลลัพธ์จากชุดข้อมูลที่มีจำนวนมากเท่าไร ยิ่งแสดงความสามารถในการเรียนรู้เชิงลึกมากเท่านั้น โดยอัลกอริทึมของวิธีการเรียนรู้เชิงลึกจะต้องใช้ ANN ซึ่งจะเหมือนกับวิธีการทำงานของระบบประสาทในสมองของมนุษย์ [8] ซึ่งปัจจุบันทั้งหน่วยงานและองค์กรต่าง ๆ ใช้วิธีการเรียนรู้ของเครื่องซึ่งเหมาะกับงานที่ต้องใช้การวิเคราะห์เป็นขั้นตอน และตัดสินใจด้วยเหตุผล เพื่อบันทึกในระบบงานคอมพิวเตอร์ให้สามารถคาดการณ์ ระบุ หรือจำแนกผลลัพธ์ที่จะเกิดขึ้น [9] เนื่องจากวิธีการเรียนรู้เชิงลึกคือซับเซ็ตย่อยของวิธีการเรียนรู้ของเครื่อง ซึ่งเป็นซอฟต์แวร์คอมพิวเตอร์ที่เลียนแบบให้เหมือนกับเครือข่ายเซลล์ประสาท (Network of Neuron) ในสมอง โดยประกอบด้วยชั้นแรกคือการนำเข้าของข้อมูล (Input Layer) และชั้นสุดท้ายคือชั้นผลลัพธ์ (Output Layer) ส่วนชั้นที่อยู่ตรงกลางเรียกว่าชั้นซ่อน (Hidden Layer) อาจมี 1 หรือมากกว่า 1 ชั้น [10] หากมีชั้นซ่อนมากกว่า 2 ชั้นซึ่งถือว่าเป็นวิธีการเรียนรู้เชิงลึก ความสามารถของวิธีการเรียนรู้เชิงลึกในอนาคตอาจจะเหนือมนุษย์ เนื่องจากสามารถเพิ่มพลังประมวลผลได้อย่างไม่จำกัด และเมื่อวิธีการเรียนรู้เชิงลึกมีหลายชั้นซ่อนทำให้สามารถคำนวณสิ่งที่ซับซ้อนได้ดี และคิดอย่างเป็นขั้นตอน ซึ่งทำให้มีประสิทธิภาพความแม่นสูง ข้อดีของวิธีการเรียนรู้เชิงลึกที่เหนือกว่าวิธีการเรียนรู้ของเครื่องทั่วไป เห็นได้ชัดเมื่อชุดข้อมูลมีขนาดใหญ่มาก เนื่องจากให้เครือข่ายสมองกลประเมินว่าควรเลือกใช้ข้อมูลตัวไหน [11] เมื่อวิธีการเรียนรู้เชิงลึกทำการจำแนกประเภทข้อมูลที่มีขนาดใหญ่มาก ปัญหาที่พบคือเวลาในการประมวลผลซึ่งต้องใช้เวลานาน และต้องใช้ข้อมูลฝึกเป็นจำนวนมากเพื่อให้ได้ประสิทธิภาพความแม่นที่สูง

ทั้งนี้ผู้วิจัยสนใจศึกษาแนวทางที่จะลดปัญหาการใช้ข้อมูลฝึกจำนวนมากเมื่อทำการจำแนกข้อมูลขนาดใหญ่มาก จึงเสนอวิธีการในการลดขนาดข้อมูลฝึก โดยเรียกว่าวิธีการที่นำเสนอ ซึ่งทำการลดขนาดข้อมูลฝึกโดยการวิเคราะห์กลุ่มด้วยวิธีเคมีนร่วมกับการตรวจหาค่าผิดปกติ (Outlier Detection) ขั้นตอนต่อมาจะนำวิธีการเรียนรู้เชิงลึกมาใช้ในการจำแนกประเภท (Classification) เพื่อให้ได้ข้อมูลฝึกที่มีขนาดลดลง แต่ยังคงมีประสิทธิภาพและมีความแม่นสูง อีกทั้งยังทำการเปรียบเทียบประสิทธิภาพความแม่นจากวิธีการเรียนรู้เชิงลึกที่ใช้การสุ่มข้อมูลฝึกขนาดเท่ากับวิธีการที่นำเสนอ และข้อมูลฝึกขนาด 80% กับ 90% ของจำนวนข้อมูลทั้งหมด

1.2. ความมุ่งหมายของการวิจัย

1.2.1 เพื่อลดขนาดข้อมูลฝึกในการจำแนกประเภทของวิธีการเรียนรู้เชิงลึกสำหรับข้อมูลขนาดใหญ่มากด้วยวิธีการที่นำเสนอ

1.2.2 เพื่อเปรียบเทียบประสิทธิภาพการจำแนกประเภทของวิธีการที่นำเสนอกับวิธีการเรียนรู้เชิงลึก

1.3 ขอบเขตของการวิจัย

1.3.1 ชุดข้อมูลขนาดใหญ่มากที่ใช้ในงานวิจัยครั้งนี้มี 2 ส่วน โดยส่วนแรกเป็นชุดข้อมูลที่สร้างขึ้น โดยสร้างคุณลักษณะ (Feature) แต่ละตัวให้มีการแจกแจงแบบต่าง ๆ 3 การแจกแจงได้แก่

1) การแจกแจงปรกติมาตรฐาน (Standard Normal Distribution)

2) การแจกแจงแบบเลขชี้กำลัง (Exponential Distribution)

3) การแจกแจงเอกรูป (Uniform Distribution)

ส่วนที่ 2 เป็นชุดข้อมูลจริง 2 ชุดข้อมูลคือ

1) ชุดข้อมูล Skin Segmentation โดยมีขนาดข้อมูล 245,057×3 (N×Feature) จากฐานเก็บข้อมูล UCI

2) ชุดข้อมูล Coil 2000 โดยมีขนาดข้อมูล 9,822×84 (N×Feature) จากฐานเก็บข้อมูล KEEL

1.3.2 วิธีการจำแนกประเภทในชุดข้อมูลขนาดใหญ่มากที่ใช้ในงานวิจัยนี้คือวิธีการเรียนรู้เชิงลึกที่มี 3 ชั้นซ่อน จำนวนโหนดในแต่ละชั้นซ่อนกำหนดเป็น 5 10 และ 20

1.3.3 ขนาดของข้อมูลฝึกที่ใช้ในงานวิจัยนี้คือข้อมูลฝึกขนาดเท่ากับวิธีการที่นำเสนอและข้อมูลฝึกขนาด 80% กับ 90% ของข้อมูลทั้งหมด

1.5 นิยามศัพท์เฉพาะ

1.5.1 ข้อมูลขนาดใหญ่มาก (Very Large-scale Data) หมายถึง มีจำนวนข้อมูล (N) × คุณลักษณะ (Feature) ≥ 500,000

1.5.2 คุณลักษณะ (Feature) หมายถึง คุณลักษณะภายในชุดข้อมูล ในทางสถิติคือตัวแปรอิสระ หรือ Independent Variable

1.5.3 เป้าหมาย (Target) หมายถึง ข้อมูลที่เป็น Class หรือผลลัพธ์ (Output) ในทางสถิติคือตัวแปรตาม หรือ Dependent Variable [13]

1.5.4 การระบุกลุ่มข้อมูล (Label) หมายถึง ตัวบ่งบอกว่าข้อมูลที่ให้ฝึกเป็นอะไร โดยจะใช้กับ Machine Learning แบบมีผู้ช่วยสอน (Supervised) เช่น ต้องสอน Machine ให้จำว่า Inputs แบบนี้ แล้วจะได้ Output แบบไหน

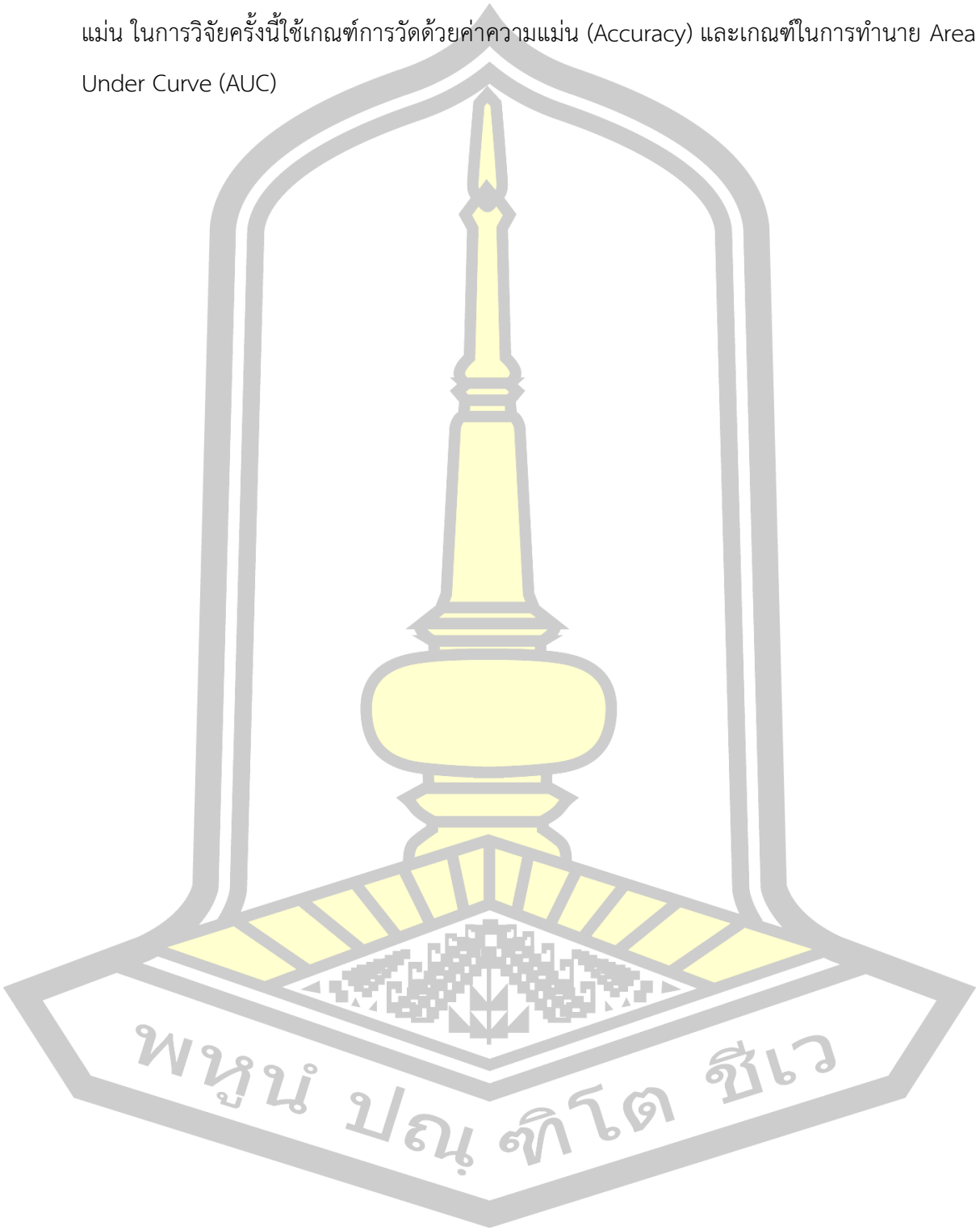
1.5.5 ข้อมูลไม่ได้ระบบุกลุ่ม (Non Label) หมายถึง ตัวบ่งบอกว่าข้อมูลที่ให้ฝึกเป็นอะไร โดยจะใช้กับ Machine Learning แบบไม่มีผู้ช่วยสอน (Unsupervised) การเรียนรู้แบบนี้ จะมีแต่ Inputs ให้ แล้วบอก Machine ว่าต้องการอะไร (เช่นการแบ่งกลุ่ม)

1.5.6 ข้อมูลฝึก (Training Data) หมายถึง ข้อมูลสำหรับฝึกตัวแบบ โดยจะฝึกให้ผลลัพธ์ออกมาเป็นไปตามชุดข้อมูลต้นฉบับ หากข้อมูลภายในชุดข้อมูลฝึกมีค่าผิดหรือค่าที่ไม่ถูกต้อง ผลลัพธ์ที่ออกมาก็จะผิด

1.5.7 ข้อมูลทดสอบ (Test Data) หมายถึง ข้อมูลสำหรับการทดสอบ โดยข้อมูลที่ใช้สำหรับทดสอบไม่ควรนำไปใช้ร่วมกับชุดข้อมูลฝึก เพราะถ้าหากใช้ร่วมกันจะทำให้การฝึกถูกต้องและแม่นยำมากไปกับข้อมูลชุดนั้น ๆ (Model Over Fitting) [14]

1.5.8 เวลาในการประมวลผล หมายถึง เวลาที่ดำเนินการตั้งแต่ทำการวิเคราะห์กลุ่มด้วยวิธีเคมีนไปจนถึงการหาประสิทธิภาพของค่าความแม่นเฉลี่ยและค่า AUC เฉลี่ย ของวิธีการเรียนรู้เชิงลึกทั้ง 27 กรณี

1.5.9 ประสิทธิภาพความแม่นของการจำแนกประเภท หมายถึง การวัดประสิทธิภาพความแม่น ในการวิจัยครั้งนี้ใช้เกณฑ์การวัดด้วยค่าความแม่น (Accuracy) และเกณฑ์ในการทำนาย Area Under Curve (AUC)



บทที่ 2
ปริทัศน์เอกสารข้อมูล

จากการจำแนกข้อมูลขนาดใหญ่มากโดยใช้การวิเคราะห์กลุ่มด้วยวิธีเคมีนและวิธีการเรียนรู้เชิงลึก ผู้วิจัยได้ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องโดยนำเสนอเนื้อหาตามลำดับดังนี้

2.1 หลักการและทฤษฎีที่เกี่ยวข้อง

2.2 งานวิจัยที่เกี่ยวข้อง

2.1 หลักการและทฤษฎีที่เกี่ยวข้อง

ในการศึกษาการจำแนกข้อมูลขนาดใหญ่มากโดยใช้การวิเคราะห์กลุ่มด้วยวิธีเคมีนและวิธี Multi-kernel SVM ของ [3] ซึ่งได้เสนอวิธีการสำหรับการจำแนกข้อมูลขนาดใหญ่มากเพื่อลดขนาดของข้อมูลฝึกและลดเวลาในการประมวลผล โดยการรวมเทคนิคการวิเคราะห์กลุ่มด้วยวิธีเคมีนและวิธี Multi-kernel SVM โดยเริ่มจากการลดขนาดของข้อมูลด้วยคุณสมบัติของวิธีเคมีนเพื่อให้ได้ข้อมูลฝึกที่เป็นตัวแทนที่ดีจากทุกส่วนของข้อมูลที่มีขนาดใหญ่มาก แล้วทำการตรวจหาค่าผิดปกติเพื่อลบข้อมูลที่ซ้ำและไกลจากข้อมูลอื่น ๆ ออก ขั้นตอนต่อมานำข้อมูลฝึกที่ได้จากการใช้คุณสมบัติของวิธีเคมีนและการตรวจหาค่าผิดปกติ เพื่อทำการจำแนกประเภทด้วยวิธี Multi-kernel SVM ผลลัพธ์ที่ได้แสดงให้เห็นว่าวิธีการเลือกข้อมูลฝึกของวิธีที่เสนอสามารถลดขนาดของข้อมูลฝึกได้ อีกทั้งยังลดเวลาที่ใช้ในการฝึกข้อมูล และสามารถรักษาประสิทธิภาพความแม่นได้ดี

เพื่อให้เกิดประโยชน์และมีประสิทธิภาพความแม่นสูงสุด ในการศึกษาครั้งนี้จึงเสนอวิธี การเรียนรู้เชิงลึกซึ่งเหมาะกับการจำแนกข้อมูลขนาดใหญ่มากและมีหลายชั้นซ่อน ทำให้สามารถคำนวณสิ่งที่ซับซ้อนมากได้ แต่ปัญหาที่ตามมาคือต้องใช้ข้อมูลฝึกเป็นจำนวนมาก ทั้งนี้จึงต้องการลดขนาดของข้อมูลฝึกด้วยการรวมเทคนิคการจัดกลุ่มของวิธีเคมีนและวิธีการเรียนรู้เชิงลึก โดยเริ่มจากการลดขนาดข้อมูลด้วยคุณสมบัติของวิธีเคมีนและการตรวจหาค่าผิดปกติ ขั้นตอนต่อมาจึงใช้วิธีการเรียนรู้เชิงลึกในการจำแนกประเภท

2.1.1 การวิเคราะห์กลุ่มด้วยวิธีเคมีน

การวิเคราะห์กลุ่มแบบไม่เป็นขั้นตอน (Nonhierarchical Cluster Analysis) หรือการวิเคราะห์กลุ่มด้วยวิธีเคมีน (K–Means Cluster Analysis) หรือการแบ่งส่วน (Partitioning) ซึ่ง

เป็นวิธีที่แตกต่างจากเทคนิคการวิเคราะห์แบบเป็นขั้นตอน (Hierarchical Cluster Analysis) เนื่องด้วยการวิเคราะห์กลุ่มแบบไม่เป็นขั้นตอนผู้วิจัยจะต้องกำหนดว่าต้องแบ่งเป็นกี่กลุ่ม เช่น Km กลุ่ม จึงเรียกวิธีนี้ว่าการวิเคราะห์กลุ่มด้วยวิธีเคมีน และชนิดของตัวแปรที่ใช้ในเทคนิคการวิเคราะห์กลุ่มด้วยวิธีเคมีนจะต้องเป็นตัวแปรเชิงปริมาณ คือเป็นสเกลอันตรภาค (Interval Scale) หรือสเกลสัดส่วน (Ration Scale) โดยไม่สามารถใช้กับข้อมูลที่อยู่ในรูปความถี่ หรือ Binary เหมือนเทคนิค Hierarchical [15]

หลักการของเทคนิคการวิเคราะห์กลุ่มด้วยวิธีเคมีนเป็นเทคนิคการจำแนก Case ออกเป็นกลุ่มย่อย จะใช้เมื่อมีจำนวน Case มาก โดยจะต้องกำหนดจำนวนกลุ่มหรือจำนวน Cluster ที่ต้องการ เช่น กำหนดให้มี Km กลุ่ม เทคนิคการวิเคราะห์กลุ่มด้วยวิธีเคมีนจะมีการทำงานหลายรอบ (Iteration) โดยในแต่ละรอบจะมีการรวม Case ให้ไปอยู่ในกลุ่มใดกลุ่มหนึ่ง โดยเลือกกลุ่มที่ Case นั้นมีระยะห่างจากค่ากลางของกลุ่มน้อยที่สุด แล้วคำนวณค่ากลางของกลุ่มใหม่ จะทำเช่นนี้จนกระทั่งค่ากลางของกลุ่มไม่เปลี่ยนแปลง หรือครบจำนวนรอบที่กำหนดไว้ ด้วยหลักการนี้จึงนำมาใช้ในการเลือกข้อมูลฝึกเพื่อให้ได้ข้อมูลฝึกที่เป็นตัวแทนที่ดีจากทุกส่วนของข้อมูล

2.1.2 หลักเกณฑ์ในการรวมกลุ่มด้วยวิธีเคมีน

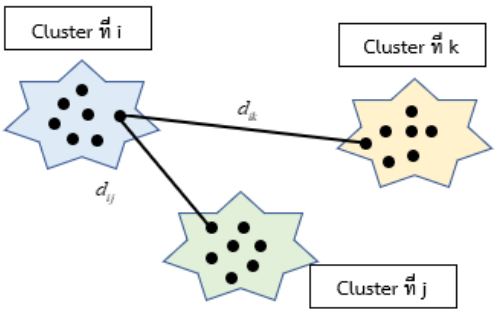
หลักเกณฑ์ในการรวมกลุ่มด้วยวิธีเคมีนซึ่งมีหลากหลายรูปแบบดังรายละเอียดต่อไปนี้ [16]

1) Nearest Neighbor หรือเรียกว่า Single Linkage วิธีการนี้ทำการรวมกลุ่มสองกลุ่มเข้าด้วยกันโดยพิจารณาจากระยะห่างที่สั้นที่สุด (ภาพที่ 1)

โดยที่ d_{ik} คือ ระยะห่างที่สั้นที่สุดระหว่างกลุ่มที่ i และกลุ่มที่ k

d_{ij} คือ ระยะห่างที่สั้นที่สุดระหว่างกลุ่มที่ i และกลุ่มที่ j

นั่นคือ ทำการรวมกลุ่มที่ i และกลุ่มที่ j เข้าเป็นกลุ่มเดียวกันเนื่องจาก $d_{ij} < d_{ik}$



ภาพที่ 1 การจำแนกประเภทด้วยหลักการของ Nearest Neighbor

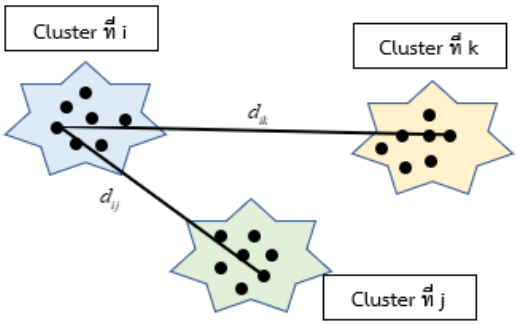
ที่มา : [16]

2) Furthest Neighbor Technique หรือเรียกว่า Complete Linkage วิธีการนี้ทำการรวมกลุ่มสองกลุ่มเข้าด้วยกันโดยพิจารณาจากระยะห่างที่ยาวที่สุด (ภาพที่ 2)

โดยที่ d_{ik} คือ ระยะห่างที่ยาวที่สุดของกลุ่มที่ i และกลุ่มที่ k

d_{ij} คือ ระยะห่างที่ยาวที่สุดของกลุ่มที่ i และกลุ่มที่ j

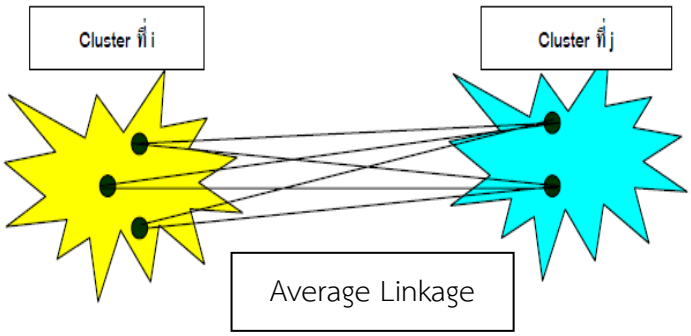
นั่นคือ $d_{ik} > d_{ij}$ จึงรวมกลุ่มที่ i และกลุ่มที่ k เข้าเป็นกลุ่มเดียวกัน



ภาพที่ 2 การจำแนกประเภทด้วยหลักการของ Furthest Neighbor Technique

ที่มา : [16]

3) Between – groups Linkage หรือเรียกว่าวิธี Average Linkage Between Groups หรือเรียกกว่า UPGMA (Unweightede Pair-Group Method Using Arithmetic Average) วิธีการนี้ทำการคำนวณหาระยะห่างเฉลี่ยของทุกคู่ของ Case โดยที่ Case หนึ่งอยู่ในกลุ่มที่ i ส่วนอีก Case หนึ่งอยู่ในกลุ่มที่ j ถ้ากลุ่มที่ i มีระยะห่างเฉลี่ยจากกลุ่มที่ j สั้นกว่าระยะห่างจากกลุ่มอื่นจะนำกลุ่มที่ i และกลุ่มที่ j รวมกันเป็นกลุ่มเดียวกัน (ภาพที่ 3)

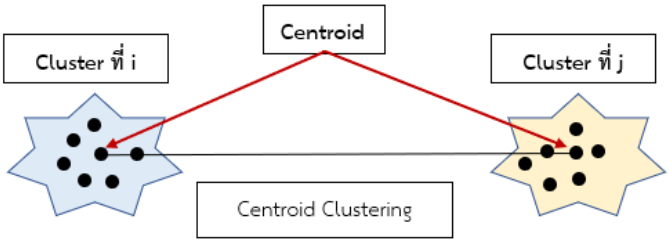


ภาพที่ 3 การจำแนกประเภทด้วยหลักการของ Between – groups Linkage

ที่มา : [16]

4) Within-group Linkage Technique วิธีการนี้ทำการรวมกลุ่มเข้าด้วยกัน ถ้าระยะห่างเฉลี่ยระหว่างทุก Case ในกลุ่มนั้น ๆ มีค่าน้อยที่สุด

5) Centroid Clustering วิธีการนี้ทำการคำนวณหาระยะห่างระหว่าง Centroid ของกลุ่มทีละคู่ ซึ่งเรียกว่าค่าเฉลี่ย หรือค่ากลางของแต่ละกลุ่มว่า Centroid เนื่องจากการจัดกลุ่ม Case จะพิจารณาจากตัวแปรหลาย ๆ ตัวพร้อม ๆ กัน จึงเรียกค่ากลางหรือค่าเฉลี่ยว่า Centroid ถ้าระยะห่างระหว่าง Centroid ของกลุ่มคู่ใดต่ำจะรวมกลุ่มคู่นั้นเข้าเป็นกลุ่มเดียวกัน (ภาพที่ 4)



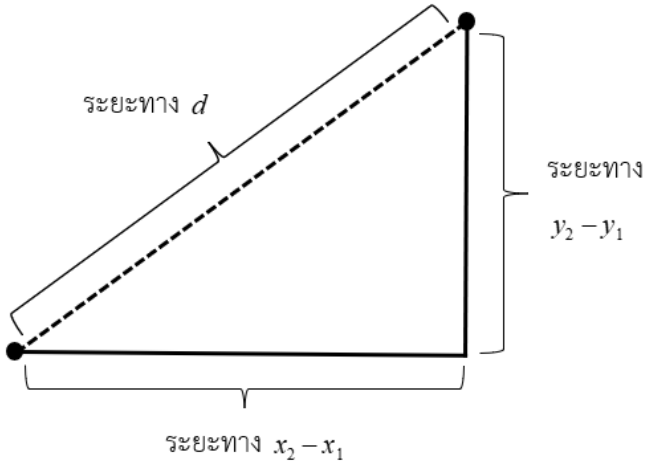
ภาพที่ 4 การจำแนกประเภทด้วยหลักการของ Centroid Clustering

ที่มา : [16]

6) Median Clustering วิธีการนี้ทำการรวมกลุ่มสองกลุ่มเข้าด้วยกัน โดยให้แต่ละกลุ่มสำคัญเท่ากัน (ให้น้ำหนักเท่ากัน) Median Clustering จะใช้ค่า Median เป็นค่ากลางของ Centroid ถ้าระยะห่างระหว่างค่า Median ของกลุ่มคู่ใดต่ำจะรวมกลุ่มคู่นั้นเข้าด้วยกัน

7) Ward's Method วิธีการนี้ทำการพิจารณาจากค่า Sum of the Squared Within-cluster Distance โดยจะรวมกลุ่มที่ทำให้ค่า Sum of square within-cluster Distance เพิ่มขึ้นน้อยที่สุด โดยค่า Square Within-cluster Distance คือค่า Square Euclidean Distance ของแต่ละ Case กับ Cluster Mean

จากหลักเกณฑ์ในการรวมกลุ่มด้วยวิธีเคมีนที่มีหลากหลายรูปแบบข้างต้น ซึ่งทำการหาระยะห่างแบบยูคลิด (Euclidean Distance) ด้วยทฤษฎีบทของปิทาโกรัสที่กล่าวไว้ว่ากำลังสองของด้านตรงข้ามมุมฉากเท่ากับผลรวมของกำลังสองของด้านประชิดมุมฉาก (ภาพที่ 5) โดยอาศัยทฤษฎีบทปิทาโกรัส เช่น การหาระยะทางจากจุดในกราฟสองจุดใด ๆ ได้ ดังนี้ [17]



ภาพที่ 5 การหาระยะทางจากจุดในกราฟสองจุดใด ๆ

เมื่อต้องการหาระยะทางจากจุด $x_1 - y_1$ ไปยังจุด $x_2 - y_2$ เมื่อกำหนดให้ d แทนระยะทาง จะได้ว่า

$$d^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2 \tag{2.1}$$

จากทฤษฎีบทปิทาโกรัสจะได้ว่า

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{2.2}$$

หาระยะทางในปริภูมิ n ใด ๆ ระยะห่างระหว่างจุด $x = (x_1, ..., x_n)$ และ $y = (y_1, ..., y_n)$ จะได้ดังสมการ (2.3)

$$d_E(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2.3}$$

โดยที่ E คือ ระยะห่างระหว่างจุดใด ๆ (Euclidean Distance)

i คือ ข้อมูลจุดใด ๆ ; n คือ จำนวนข้อมูลทั้งหมด

2.1.3 การตรวจหาค่าผิดปกติ (Outlier Detection)

ตรวจหาค่าผิดปกติเพื่อลดขนาดของข้อมูลฝึก ในขั้นแรกรวบรวมข้อมูลที่ผ่านคุณสมบัติของวิธีเคมีนเพื่อทำการลบข้อมูลที่มีค่าซ้ำออก เพื่อไม่ให้เกิดความซ้ำซ้อนของข้อมูล และลดขนาดข้อมูลลง ค่าของข้อมูลที่ผิดปกติมีผลกระทบด้านลบต่อการสร้างตัวจำแนกประเภท จึงใช้วิธีการตรวจหาค่าผิดปกติที่เสนอใน [18] เพื่อหาข้อมูลที่เป็นค่าผิดปกติ จะคำนวณสถิติทดสอบ คอลโมโกรอฟ-สมีร์นอฟ (Kolmogorov-Sminov) ระหว่างข้อมูลที่ j ไปยังข้อมูลจุดอื่น ๆ ที่รวบรวมมาโดยสูตรดังนี้

$$KS(p_j - p_i) = sup_x \left| Fp_j(x) - Fp_i(x) \right| \tag{2.4}$$

โดยที่ Fp_j คือ ระยะทางจากจุด j ไปยังจุดอื่น ๆ ที่รวบรวม

Fp_i คือ ระยะทางจากจุด i ไปยังจุดอื่น ๆ ที่รวบรวม

ค่าเฉลี่ยของสถิติทดสอบคอลโมโกรอฟ-สมีร์นอฟจะใช้ในการคำนวณสถิติทดสอบ KSE ซึ่งจะคำนวณการตรวจหาค่าผิดปกติโดยสูตร ดังนี้

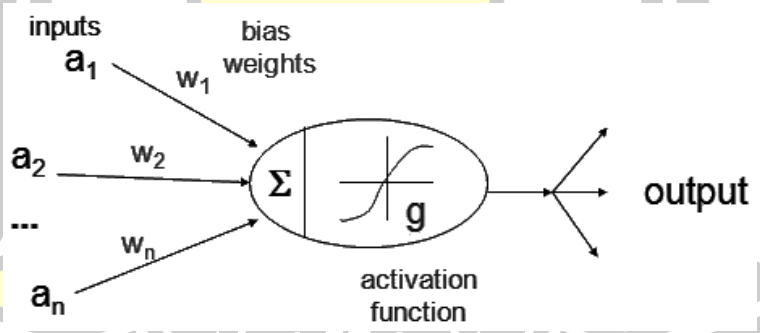
$$KSE(p_j) = \frac{1}{n-1} \sum_{i=1}^{n} KS(p_j - p_i) \tag{2.5}$$

ถ้าค่าเฉลี่ยสถิติ KS ที่เรียกว่าสถิติ KSE มีขนาดใหญ่กว่าเกณฑ์สำหรับทดสอบข้อมูลจะถือว่าตัวอย่างนั้นเป็นค่าผิดปกติและถูกลบออกทันที เนื่องจากยากที่จะหาเกณฑ์ที่เหมาะสม ดังนั้นจึงลบข้อมูลที่มีค่า KSE ที่สูงสุดออก และทำซ้ำขั้นตอนนี้หลาย ๆ ครั้ง จำนวนครั้งของการทำซ้ำจะถูกกำหนดว่าเป็นจำนวนครั้งในการตรวจสอบค่าผิดปกติ (ROT) หลังจากนั้นจะได้ชุดข้อมูลฝึกที่ลดลง [3]

2.1.4 วิธีการเรียนรู้เชิงลึก (Deep Learning)

วิธีการเรียนรู้เชิงลึกคือซอฟแวร์คอมพิวเตอร์ที่เลียนแบบการทำงานของระบบโครงข่ายประสาทในสมองของมนุษย์ซึ่งเป็นรูปแบบย่อยของวิธีการเรียนรู้ของเครื่อง [7] โดยวิธีการเรียนรู้เชิงลึกใช้หลักการจาก ANN คือการจำลองวิธีการทำงานของสมองให้คอมพิวเตอร์รู้จักคิด และจดจำคล้ายโครงข่ายประสาทของมนุษย์ แต่การทำงานของ ANN จะไม่ซับซ้อนเท่าระบบประสาทมนุษย์ โดยวิธีการเรียนรู้เชิงลึกคือการนำแนวคิดของ ANN มาใช้ในระดับที่ลึกและมีความซับซ้อนมากกว่าหรือมีชั้นที่เยอะมากกว่า และสามารถหาวิธีการแก้ปัญหาได้เร็วและแม่นยำกว่า ANN ซึ่งสามารถรู้ว่าเป็นวิธีการเรียนรู้เชิงลึกอย่างชัดเจนได้ยาก [19] และในปี 2018 [11] กล่าวว่าหากมีชั้นซ่อนมากกว่า 2 ชั้นถือว่าเป็นวิธีการเรียนรู้เชิงลึกแล้ว ทั้งนี้ขึ้นอยู่กับจุดมุ่งหมายของสิ่งที่ต้องการหา เนื่องด้วย ANN ที่สามารถใช้ในการทำนาย (Prediction) หรือจำแนกประเภทได้แล้ว ยังถูกนำไปต่อยอดเป็น Deep Learning (DL) โครงข่ายประสาทเทียมแบบวนซ้ำ (Recurrent Neural Network หรือ RNN) และโครงข่ายประสาทเทียมแบบสังวัตนาการ (Convolutional Neural Network หรือ CNN) จากการทำงานของระบบโครงข่ายประสาทเทียมในสมองมนุษย์และถือเป็นส่วนหนึ่งของวิธีการเรียนรู้ของเครื่อง [20]

ANN ซึ่งมาจากการจำลองระบบประสาทของสิ่งมีชีวิตขึ้นมาให้อยู่บนสิ่งที่คอมพิวเตอร์สามารถจำลองและคำนวณออกมาได้ (ภาพที่ 6)



ภาพที่ 6 ประสาทเทียมที่ถูกแปลงเป็นฟังก์ชันทางคณิตศาสตร์

ที่มา : [20]

สามารถคำนวณได้ดังนี้

$$Output = F\left(\sum_{i=1}^{n} x_i w_i\right) \quad (2.6)$$

โดยที่ F คือ ฟังก์ชันที่รับผลรวมการประมวลผลทั้งหมด จากทุก Input (Activation Function)

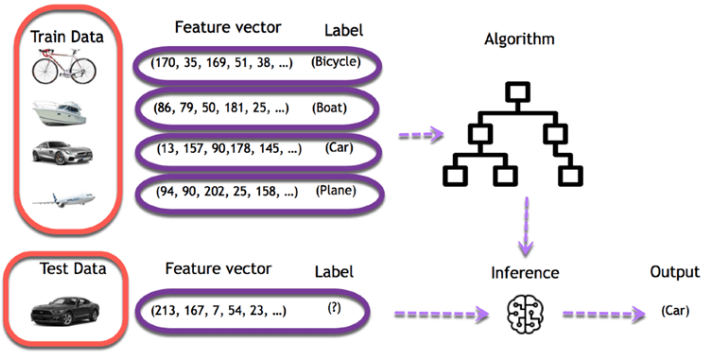
x_i คือ ข้อมูลตัวที่ i

w_i คือ น้ำหนักของข้อมูลตัวที่ i

วิธีการเรียนรู้ของเครื่องเป็นหนึ่งในศาสตร์ความรู้ของปัญญาประดิษฐ์ หรือ Artificial Intelligence : AI ซึ่งหมายถึงความสามารถของระบบไอทีในการหาทางแก้ปัญหาด้วยตนเอง โดยการจดจำรูปแบบในฐานข้อมูล กล่าวคือวิธีการเรียนรู้ของเครื่องช่วยให้ระบบไอทีรู้จักรูปแบบพื้นฐานของอัลกอริทึมและชุดข้อมูลที่มีอยู่ เพื่อพัฒนากระบวนการแก้ปัญหาที่เหมาะสม ดังนั้นในวิธีการเรียนรู้ของเครื่องจึงถูกสร้างขึ้นบนพื้นฐานของประสบการณ์เพื่อให้ซอฟต์แวร์สามารถสร้างการแก้ปัญหาได้อย่างอิสระ ซึ่งจำเป็นต้องมีการกระทำก่อนหน้านี้ของมนุษย์ นั่นคือการป้อนชุดอัลกอริทึมและข้อมูลที่จำเป็นลงในระบบล่วงหน้า รวมถึงกฎการวิเคราะห์ตามลำดับเพื่อการจดจำรูปแบบในสต็อคข้อมูล เมื่อทั้งสองขั้นตอนนี้เสร็จสิ้นระบบจึงจะสามารถแสดงผลการทำงานของวิธีการเรียนรู้ของเครื่องได้

วิธีการเรียนรู้ของเครื่องจะทำงานคล้ายกับการเรียนรู้ของมนุษย์ เช่น หากแสดงภาพวัตถุประเภทต่าง ๆ ให้เด็กคนหนึ่งดู พวกเขาก็จะสามารถเรียนรู้ที่จะระบุและแยกความแตกต่างในภาพเหล่านั้นได้ วิธีการเรียนรู้ของเครื่องทำงานในลักษณะเดียวกันด้วยการป้อนชุดข้อมูลและชุดคำสั่งต่าง ๆ คอมพิวเตอร์จะถูกใช้งานเพื่อเรียนรู้และจำแนกประเภทของแต่ละชนิดในวัตถุนั้น เช่น การทำนายประเภทของวัตถุในกระบวนการของวิธีการเรียนรู้ของเครื่อง (ภาพที่ 7) ด้วยเหตุนี้ซอฟท์แวร์จึงถูกทดแทนด้วยข้อมูลและการฝึก เช่น โปรแกรมเมอร์สามารถบอกระบบได้ว่าวัตถุใดเป็นมนุษย์ (Human) และอีกวัตถุหนึ่งไม่ใช่มนุษย์ (Non Human) ซอฟต์แวร์จะได้รับการป้อนข้อมูลตอบรับ (Feedback) อย่างต่อเนื่องจากโปรแกรมเมอร์ สัญญาณการตอบรับเหล่านี้จะถูกใช้โดยอัลกอริทึมเพื่อปรับและเพิ่มประสิทธิภาพตัวแบบ ด้วยชุดข้อมูลใหม่ที่ป้อนเข้าไปในระบบ รูปแบบจะได้รับการปรับให้เหมาะสมเพื่อให้สามารถแยกแยะระหว่าง Humans กับ Non-humans ได้อย่างชัดเจนในตอนท้าย วิธีการเรียนรู้ของเครื่องช่วยให้มนุษย์ทำงานได้อย่างสร้างสรรค์และมีประสิทธิภาพมากขึ้น

สามารถมอบหมายงานที่ค่อนข้างซับซ้อนหรือซ้ำซากให้กับคอมพิวเตอร์ผ่านวิธีการเรียนรู้ของเครื่องได้ เริ่มจากการสแกน บันทึก และจัดเก็บเอกสาร เช่น ใบแจ้งหนี้เพื่อจัดระเบียบและแก้ไขภาพ นอกเหนือจากงานง่าย ๆ เหล่านี้ เครื่องจักรเรียนรู้ด้วยตนเอง (Self-learning Machine) ยังสามารถทำงานที่ซับซ้อนได้ เช่น การจดจำรูปแบบที่ผิดพลาดต่าง ๆ (Error Patterns) ได้ด้วย นี่เป็นข้อได้เปรียบที่สำคัญโดยเฉพาะอย่างยิ่งในพื้นที่ต่าง ๆ เช่น อุตสาหกรรมการผลิต ซึ่งเป็นอุตสาหกรรมที่ต้องพึ่งพาการผลิตอย่างต่อเนื่องและปราศจากข้อผิดพลาด แม้แต่ในเรื่องที่ผู้เชี่ยวชาญก็ไม่อาจมั่นใจได้ว่าเกิดข้อผิดพลาดในสายการผลิตได้อย่างไร แต่วิธีการเรียนรู้ของเครื่องเสนอความเป็นไปได้ในการระบุข้อผิดพลาดได้อย่างรวดเร็ว ซึ่งช่วยประหยัดเวลาและค่าใช้จ่าย [21]



ภาพที่ 7 การทำนายประเภทของวัตถุในกระบวนการของวิธีการเรียนรู้ของเครื่อง

ที่มา : [7]

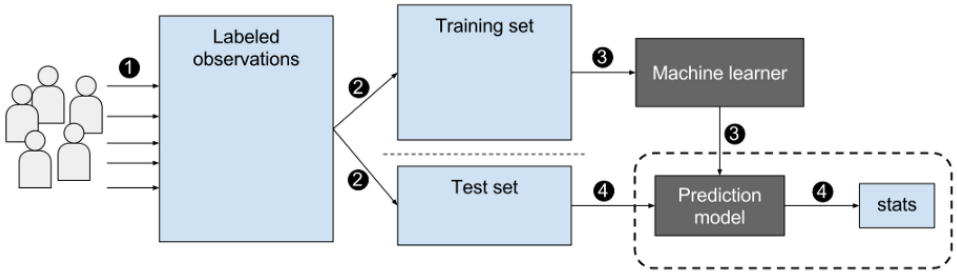
วิธีการเรียนรู้ของเครื่องประกอบไปด้วย 2 แบบใหญ่ ๆ คือ แบบมีผู้ช่วยสอน (Supervised) และแบบไม่มีผู้ช่วยสอน (Unsupervised)

1. แบบมีผู้ช่วยสอน (Supervised) ในแบบนี้จะมีสิ่งที่เรียกว่าเฉลย (Label หรือ Class) มีหน้าที่จำแนกประเภทของข้อมูลนั้น ๆ (Category) หรือบ่งบอกถึงปริมาณก็ได้ โดยสามารถแบ่งออกมาได้อีก 2 ประเภทคือ การจำแนกประเภท (Classification) หรือ การวิเคราะห์การถดถอย (Regression)

การจำแนกประเภทเป็นการจำแนกข้อมูลออกเป็นประเภทต่าง ๆ ตามที่ Label ได้กำหนดไว้ ซึ่งวิธีการเรียนรู้ของเครื่องทำการจำแนกประเภทโดยให้คำตอบเป็น Label หรือ Class เท่านั้น ไม่สามารถให้คำตอบที่นอกเหนือจาก Label ในชุดข้อมูลการฝึกหรือออกมาเป็นตัวเลขที่ผ่าน

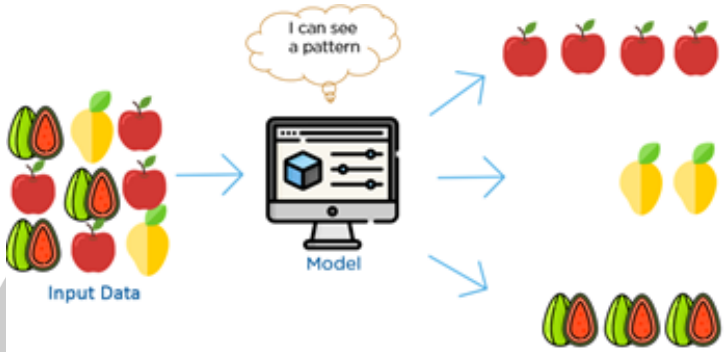
การคำนวณได้นั้นเอง (ภาพที่ 8) โดยตัวแบบของวิธีการเรียนรู้ของเครื่องที่เหมาะสำหรับงานด้านการจำแนกประเภท ได้แก่ 1) KNN 2) SVM 3) Logistic Regression 4) Decision Tree เป็นต้น

การวิเคราะห์การถดถอยคือการนำ Input เข้าไปฝึกฝนและให้คำตอบออกมาเป็นตัวเลขเท่านั้น คำตอบไม่สามารถออกมาเป็น Label หรือ Class ได้ โดยตัวแบบของวิธีการเรียนรู้ของเครื่องที่เหมาะสำหรับงานการวิเคราะห์การถดถอย ได้แก่ 1) Linear Regression 2) Ridge Regression 3) Lasso 4) Elastic Net 5) SGD เป็นต้น



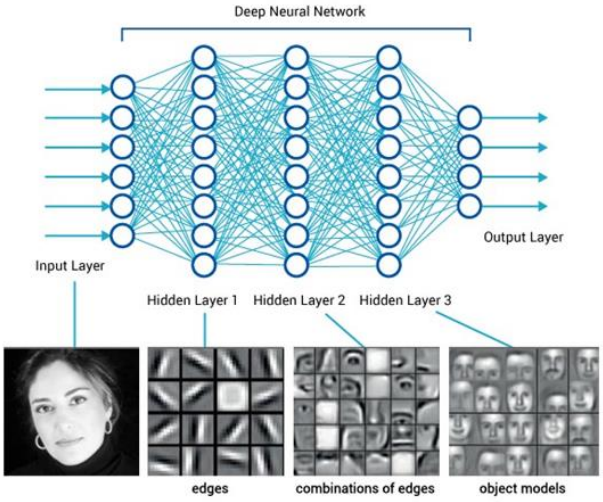
ภาพที่ 8 จำแนกประเภทของข้อมูล โดยมีผู้ช่วยสอน (Supervised)
ที่มา : [14]

2. แบบไม่มีผู้ช่วยสอน (Unsupervised) จะต่างจาก Supervised โดยสิ้นเชิงโดยวิธีนี้จะเน้นไปที่การวิเคราะห์ข้อมูลมากกว่า เช่น การหารูปแบบของข้อมูลเพื่อทำการจัดกลุ่มของข้อมูล (Clustering) หรือจะเป็นการลดมิติของข้อมูล (Dimension Reduction) เพื่อหาคุณลักษณะของข้อมูล (ภาพที่ 9)



ภาพที่ 9 จำแนกประเภทของข้อมูล โดยไม่ต้องมีผู้ช่วยสอน (Unsupervised)
ที่มา : [14]

อัลกอริทึมของวิธีการเรียนรู้เชิงลึกถูกสร้างขึ้นจากการนำเอา ANN หลาย ๆ ชั้นมาต่อกัน โดยชั้นแรกจะทำหน้าที่ในการรับข้อมูล (Input Layer) และชั้นสุดท้ายจะทำหน้าที่ส่งผลลัพธ์การประมวลผลออกมา (Output Layer) ส่วนชั้นที่อยู่ระว่างกลางจะถูกเรียกว่าชั้นซ่อน (Hidden Layer) วิธีการเรียนรู้เชิงลึกมีที่มาจากการที่ใช้ชั้นของ ANN หลาย ๆ อันมาต่อกัน เนื่องจากชั้นเหล่านี้เป็นโครงสร้างที่ถูกจัดเก็บแบบเป็นกองซ้อน ๆ (Stack) จึงเปรียบได้ว่าชั้นจำนวนเยอะมาก ๆ จะทำให้มีโครงสร้างที่ลึกมากยิ่งขึ้น (ภาพที่ 10)



ภาพที่ 10 วิธีการเรียนรู้เชิงลึกที่มีหลายชั้นซ่อน (Hidden Layer)

ที่มา : [11]

โดยชั้นซ่อนของแต่ละชั้นจะเปรียบเสมือนว่าประกอบด้วยเซลล์ประสาทจำนวนมาก ซึ่งมีหน้าที่ในการประมวลผล รับข้อมูลจากชั้นที่อยู่เหนือกว่า และส่งข้อมูลที่ประมวลผลเสร็จแล้วไปยัง ชั้นที่อยู่ล่างกว่า ข้อดีของการส่งข้อมูลแบบนี้ก็คือชั้นแต่ละชั้นสามารถที่จะมีค่าถ่วงน้ำหนัก (Weight) ค่าความเอนเอียงของข้อมูล (Bias) และ วิธีการประมวลผลทางคณิตศาสตร์ (Activation Function) ที่เป็นอิสระต่อกันได้ และเมื่อทำการป้อนข้อมูลให้กับตัวแบบมาก ๆ ชั้นแต่ละชั้นก็จะสามารถสกัด คุณลักษณะที่มีความซับซ้อนมากยิ่งขึ้น ตัวแบบที่ใช้วิธีการเรียนรู้เชิงลึกจะให้ความแม่น (Accuracy) ที่สูงในหลาย ๆ ปัญหา ตั้งแต่การตรวจจับวัตถุ (Object Detection) ไปจนถึงการรู้จำเสียงพูด (Speech Recognition) โดยที่ไม่จำเป็นต้องให้ความรู้พื้นฐานใด ๆ ไว้ล่วงหน้า เพียงแค่ให้ข้อมูลตัวอย่างก็จะทำการเรียนรู้จากข้อมูลและสังเคราะห์เป็นองค์ความรู้ออกมาได้อย่างอัตโนมัติ เช่น การใช้วิธีการเรียนรู้เชิงลึกในวงการเกม ไม่จำเป็นต้องบอกว่าเล่นยังไง เพียงแค่ให้เรียนรู้จากผู้เล่นที่เก่ง ๆ เป็นจำนวนมาก วิธีการเรียนรู้เชิงลึกก็จะเรียนรู้วิธีการเล่นเกมได้อย่างอัตโนมัติ

ปัจจุบันได้นำวิธีการเรียนรู้เชิงลึกมาประยุกต์ใช้อย่างแพร่หลาย เช่น รถยนต์ไร้คนขับ (Driverless Car) สมาร์ทโฟน Search Engine ของ Google เครื่องจับเท็จ (Fraud Detection) โทรทัศน์ และอื่น ๆ อีกมากมาย วิธีการเรียนรู้เชิงลึกถือเป็นเครื่องมือที่ทรงพลังในการ

ทำนายผลลัพธ์ต่าง ๆ อีกทั้งยังสามารถหารูปแบบหรือสังเคราะห์ข้อมูลได้เหนือกว่าองค์ความรู้เดิมที่มีอยู่ (Unsupervised Learning) ข้อมูลขนาดใหญ่ หรือ Big Data ก็เปรียบเสมือนเชื้อเพลิงของวิธีการเรียนรู้เชิงลึก การผสมผสานระหว่างทั้งสองอย่างเรียกได้ว่าสามารถทำให้มนุษยชาติก้าวไปขั้นอีกขั้น ไม่ว่าจะเป็นด้านของผลผลิต การขาย การบริหาร และนวัตกรรม อีกทั้งยังทำงานมีประสิทธิภาพได้ดีกว่าวิธีการดั้งเดิม กล่าวคืออัลกอริทึมของวิธีการเรียนรู้เชิงลึกให้ความแม่นมากว่าอัลกอริทึมของวิธีการเรียนรู้ของเครื่อง ในด้านของการจำแนกภาพ (Image Classification) ด้านของการรู้จำใบหน้า (Face Recognition) และการรู้จำเสียง (Voice Recognition) วิธีการเรียนรู้เชิงลึกถูกนำมาประยุกต์ใช้ในหลากหลายวงการ ตั้งแต่การเงิน (Finance) ไปถึงการตลาด (Marketing) ห่วงโซ่อุปทาน (Supply Chain) และบริษัทยักษ์ใหญ่จะเป็นบริษัทแนวหน้าที่สามารถใช้วิธีการเรียนรู้เชิงลึกได้อย่างมีประสิทธิภาพ เพราะบริษัทเหล่านี้มีข้อมูลมากพอที่จะทำการฝึกข้อมูล [7] ในปี 2561 [22] กล่าวว่ากรณี Overfitting และ Underfitting ซึ่งเป็นข้อผิดพลาดในการสร้างวิธีการเรียนรู้เชิงลึกที่อาจเกิดขึ้นได้จากการจำแนกประเภท ซึ่ง Overfitting คือตัวแบบที่ตอบสนองต่อการรบกวน (Noise) จำนวนมาก จนเริ่มเรียนรู้จากการรบกวนและรายละเอียดของข้อมูลที่ไม่ถูกต้อง ทำให้ตัวแบบไม่เหมาะสำหรับการทำนายข้อมูล เพราะมีรายละเอียดและการรบกวนมาก กรณีนี้ตัวแบบมีค่าความแปรปรวนของข้อมูลสูง (High Variance) เช่น ตัวแบบที่ทำการฝึกแล้วนำชุดข้อมูลที่ฝึกวัดประสิทธิภาพความแม่นได้ 99% แต่เมื่อนำชุดข้อมูลทดสอบ (Test) วัดประสิทธิภาพความแม่นได้เพียง 60% นั่นคือ Overfitting ส่วน Underfitting คือตัวแบบที่ไม่สามารถทำงานได้ จากการที่ไม่สามารถจับแนวโน้มของข้อมูลได้ อันเนื่องมาจากตัวแบบไม่เหมาะสมหรือข้อมูลมีจำนวนน้อยไป กรณีนี้ตัวแบบมีค่าความเอนเอียงสูง (High Bias) เช่น หากนำชุดข้อมูลที่ฝึกวัดประสิทธิภาพความแม่นจะได้ค่าความแม่นต่ำ และเมื่อนำชุดข้อมูลทดสอบวัดประสิทธิภาพความแม่นจะได้ค่าความแม่นต่ำเช่นกัน สามารถแก้ปัญหาโดยเพิ่มข้อมูลมากขึ้น และลดคุณลักษณะลง (Features)

2.1.5 การสุ่มข้อมูลด้วยวิธี K-fold Cross Validation

ในการสุ่มข้อมูลฝึกของวิธีการเรียนรู้เชิงลึกขนาด 80% กับ 90% จากข้อมูลทั้งหมด โดยทำการสุ่มด้วยวิธี K-fold [33] กล่าวว่าวิธีการของ K-fold คล้ายกับ Leave one out ต่างกันที่เลือกจากครั้งละตัวแทนเป็นเลือกครั้งละ N/K ตัวแทน (ซึ่ง K เป็นค่าคงที่ที่เลือกแต่ละตัว) และทำ K ครั้ง ด้วยวิธีการนี้ไม่ว่าข้อมูลจะมีมากน้อยเพียงใด ในการเรียนรู้ของเครื่องจะทำการฝึกฝน K ครั้งเท่านั้น ซึ่งค่า K ที่นิยมและถือว่าเป็นมาตรฐานก็คือ K=10 เพราะจะเหลือข้อมูลไว้สำหรับ

ฝึกถึง 90% ในแต่ละรอบ (Training Data) และมี 10% ไว้สำหรับทดสอบ (Testing Data) ทั้งนี้จึงเรียกว่าวิธี K-fold Cross Validation

เมื่อ K = 10 หมายถึงการแบ่งข้อมูลออกเป็น 10 ส่วน ในแต่ละส่วนมีจำนวนข้อมูลเท่า ๆ กัน โดยสร้างแบบจำลองจากข้อมูล 9 ชุด อีก 1 ชุดที่เหลือใช้สำหรับทดสอบ วนซ้ำการทำงานนี้จนกระทั่งข้อมูลทุกส่วนถูกใช้เป็นชุดทดสอบ [34] ดังภาพที่ 11

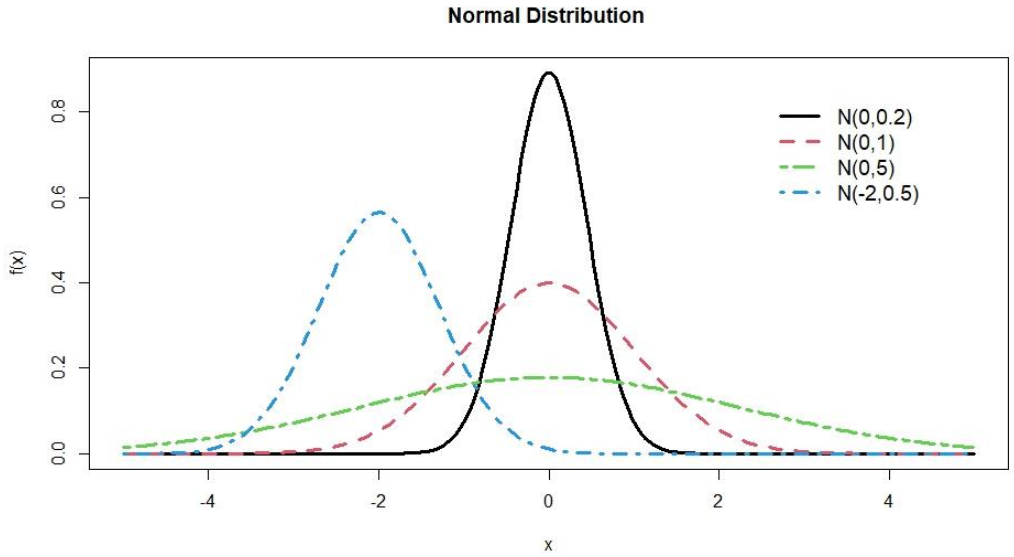
Training	Training	Training	Training	Training	Training	Training	Training	Training	Training
2	1	1	1	1	1	1	1	1	1
3	3	2	2	2	2	2	2	2	2
4	4	4	3	3	3	3	3	3	3
5	5	5	5	4	4	4	4	4	4
6	6	6	6	6	5	5	5	5	5
7	7	7	7	7	7	6	6	6	6
8	8	8	8	8	8	8	7	7	7
9	9	9	9	9	9	9	9	8	8
10	10	10	10	10	10	10	10	10	9
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
Model	Model	Model	Model	Model	Model	Model	Model	Model	Model
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
1	2	3	4	5	6	7	8	9	10
Testing	Testing	Testing	Testing	Testing	Testing	Testing	Testing	Testing	Testing
รอบที่ 1	รอบที่ 2	รอบที่ 3	รอบที่ 4	รอบที่ 5	รอบที่ 6	รอบที่ 7	รอบที่ 8	รอบที่ 9	รอบที่10

ภาพที่ 11 การแบ่งข้อมูลแบบ K-fold Cross Validation กรณี K = 10

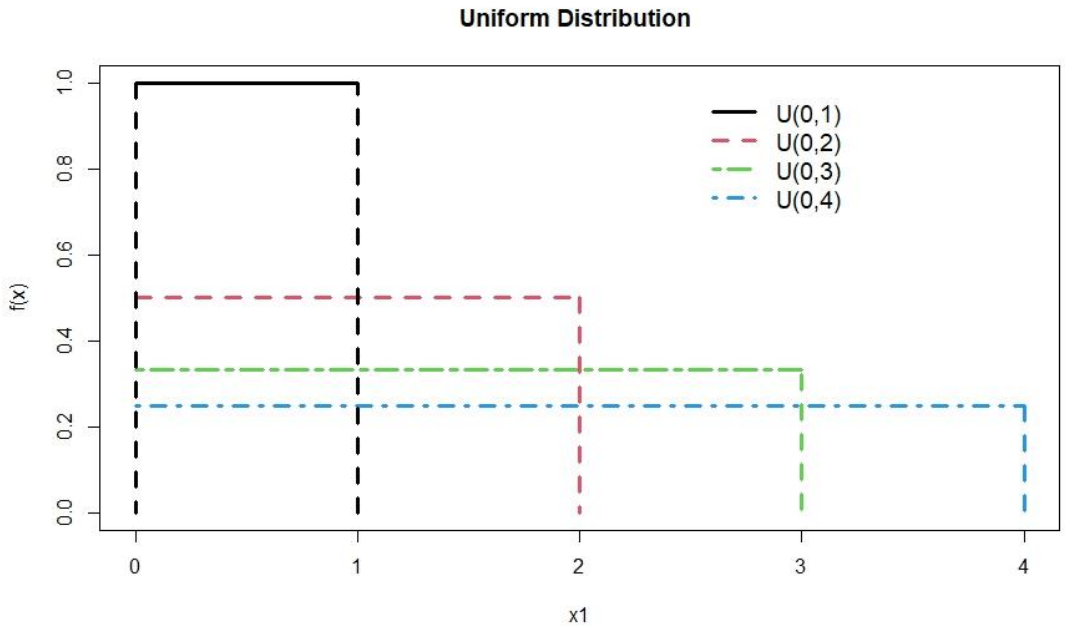
2.1.6 การแจกแจงของข้อมูล

การแจกแจงของข้อมูลที่นำมาศึกษาในครั้งนี้มี 3 การแจกแจงได้แก่ 1) การแจกแจงปรกติมาตรฐาน (Standard Normal Distribution) 2) การแจกแจงแบบเลขชี้กำลัง (Exponential Distribution) 3) การแจกแจงเอกรูป (Uniform Distribution) ซึ่งการแจกแจงปรกติมาตรฐานและการแจกแจงแบบเอกรูปเป็นตัวแทนของการแจกแจงแบบสมมาตร โดยที่การแจกแจงปรกติมาตรฐาน

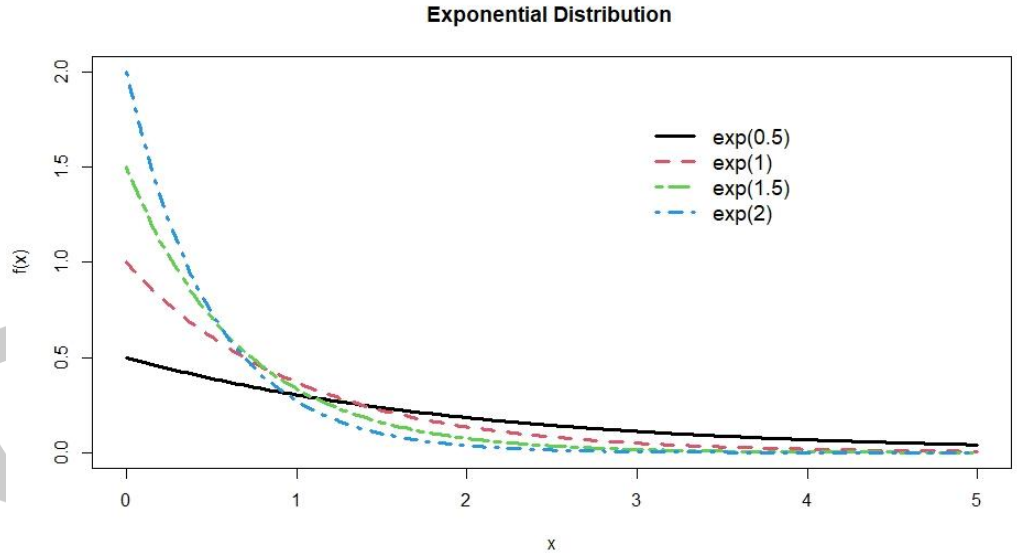
มีลักษณะเป็นรูประฆังคว่ำ โดยข้อมูลส่วนใหญ่จะอยู่ที่ค่า $\mu = 0$ -3 ถึง 3 ดังภาพที่ 12 ในขณะที่การแจกแจงแบบเอกรูปข้อมูลมีลักษณะที่กระจายเป็นรูปสี่เหลี่ยมผืนผ้า ดังภาพที่ 13 โดยการแจกแจงแบบเอกรูปที่ศึกษา ข้อมูลจะมีค่าอยู่ระหว่าง 0 ถึง 1 จะเห็นได้ว่าถึงแม้ว่าพารามิเตอร์เปลี่ยนไปแต่ลักษณะของข้อมูลยังคงสมมาตร สำหรับการแจกแจงแบบเลขชี้กำลังเป็นตัวแทนของข้อมูลที่มีการแจกแจงแบบเบ้ขวา ดังภาพที่ 14



ภาพที่ 12 เปรียบเทียบการแจกแจงปรกติ (Standard Normal Distribution)



ภาพที่ 13 การแจกแจงเอกรูป (Uniform Distribution)



ภาพที่ 14 การแจกแจงแบบเลขชี้กำลัง (Exponential Distribution)

2.2 งานวิจัยที่เกี่ยวข้อง

2.2.1 งานวิจัยในประเทศไทย

การประยุกต์วิธีการเรียนรู้โดยใช้โครงข่ายประสาทเทียมหลายโครงข่ายบนข้อมูลขนาดใหญ่ การเรียนรู้และวิเคราะห์ข้อมูลขนาดใหญ่ด้วยวิธีการเรียนรู้โดยใช้โครงข่ายประสาทเทียมเป็นเรื่องที่สำคัญมากในการทำเหมืองข้อมูล แต่มักจะเกิดปัญหาในด้านเวลาที่ใช้ในการเรียนรู้ ในงานวิจัยนี้ได้เสนอวิธีการเรียนรู้โดยการใช้วิธีการเรียนรู้โดยใช้โครงข่ายประสาทเทียมหลายโครงข่ายทำการเรียนรู้บนชุดข้อมูลฝึก ที่ถูกแบ่งย่อยและสุ่มเลือกมาจากชุดข้อมูลทั้งหมด จากนั้นจึงทำการรวมโหนดในชั้นซ่อนจากวิธีการเรียนรู้โดยใช้โครงข่ายประสาทเทียมแต่ละอัน เพื่อหาค่าน้ำหนักประจำโหนดในชั้นซ่อนใหม่ที่เหมาะสมสำหรับข้อมูลทั้งชุด ผู้วิจัยได้ทำการพัฒนาอัลกอริทึมการรวมโหนด โดยประยุกต์จากการหาระยะทางแบบยุคลิดเพื่อระบุความใกล้เคียงกันของโหนด เพื่อที่จะรวมค่าน้ำหนักของโหนดที่ใกล้เคียงเข้าไว้ด้วยกัน ผลการทดลองพบว่าวิธีการที่นำเสนอสามารถลดเวลาในการเรียนรู้ลงได้อย่างมาก และยังคงรักษาประสิทธิภาพความแม่นได้เหมือนกับการใช้ข้อมูลทั้งหมด [4]

โครงข่ายประสาทเทียมที่ใช้ข้อมูลจากข้อมูลความเร่งในการระบุตัวคนขับรถโดยใช้ฮิสโทแกรม โดยเสนอระบบการระบุตัวคนขับรถซึ่งใช้เพียงข้อมูลจากตัวรับรู้ความเร่ง โดยมีการใช้ฮิสโทแกรมของความเร่งเป็นข้อมูลนำเข้าสู่โครงข่ายประสาทเทียม สถาปัตยกรรมระบบที่นำเสนอในงานวิจัยนี้สามารถใช้เป็นแนวทางในการพัฒนาระบบระบุตัวคนขับรถอื่นในอนาคต ผลการทดสอบประสิทธิภาพของระบบนี้สามารถระบุตัวคนขับรถได้แม่นยำสูงสุดถึง 99% นอกจากนี้ในงานวิจัยยังได้ทดสอบประสิทธิภาพในหลายแง่มุมซึ่งที่ผ่านมามีหลายงานวิจัยที่มองข้ามบางแง่มุมนี้ไป ดังนั้นการวัดผลในงานวิจัยนี้จึงสามารถใช้เป็นแนวทางในการวัดผลการระบุตัวคนขับรถอื่นในอนาคตได้เช่นกัน [23]

การใช้โครงข่ายประสาทเทียมเพื่อนำเสนอกระบวนการค้นคืนรูปภาพลายผ้าไหมที่มีกลุ่ม ตัวอย่างน้อย โดยวิธีการตรวจหาจุดสนใจภาพร่วมกับการสกัดคุณลักษณะพิเศษเฉพาะพื้นที่ด้วยวิธี Scale-Invariant Feature Transform (SIFT) และวิธีการสุ่มข้อมูลเพื่อหาความสอดคล้องของกลุ่มตัวอย่าง (Random Sample consensus: RANSAC) เปรียบเทียบระยะห่างของภาพลายผ้าไหมจากการวัดระยะทาง (Distance Measurement) ระหว่างจุดสนใจภาพลายผ้าไหมในแต่ละรอบเปรียบเทียบประสิทธิภาพระหว่างวิธีการหาคุณลักษณะพิเศษเฉพาะพื้นที่ และโครงข่ายประสาทเทียม

แบบคอนโวลูชัน (CNN) สำหรับการค้นคืนรูปภาพลายผ้าไหมไทย วิธีการหาคุณลักษณะพิเศษเฉพาะพื้นที่ถูกนำมาเพื่อเปรียบเทียบในการสร้างข้อมูลลักษณะพิเศษ ประกอบด้วยวิธี Histogram of Oriented Gradients (HOG) และวิธี SIFT ดังนั้นข้อมูลลักษณะพิเศษจะถูกส่งไปเพื่อคำนวณร่วมกับวิธี K-Nearest Neighbor (KNN) และวิธี Support Vector Machine (SVM) โครงสร้างของวิธี CNN ที่ใช้ในการทดลองประกอบด้วยโครงสร้างแบบ LeNet-5 และ AlexNet จากการทดลองพบว่าวิธีการตรวจหาจุดสนใจที่นำเสนอมีอัตราการค้นคืนสูงสุดโดยเฉลี่ยเท่ากับ 95.69% ใน Top-1 และวิธีการหาคุณลักษณะพิเศษเฉพาะพื้นที่เมื่อนำไปคำนวณร่วมกับวิธี KNN และวิธี SVM มีประสิทธิภาพสูงกว่าวิธี CNN [24]

ในการใช้เทคนิคการเรียนรู้เชิงลึกผ่านชุดโครงข่ายประสาทเทียม เพื่อหาแบบจำลองทำนายผลคำตัดสินและประเด็นในคดีอาญาที่เรียนรู้จากคำพิพากษาศาลฎีกาไทย แบบจำลองนี้สร้างตัวแทนข้อความด้วยโครงข่ายประตูวกกลับสองทิศทางร่วมด้วยกลไกจุดสนใจ ก่อนนำตัวแทนข้อความนั้นไปทำนายผลคำตัดสินและประเด็นในคดีอาญาด้วยโครงข่ายประสาทเทียมแบบโมดูลซึ่งจำลองโครงสร้างความรับผิดทางอาญาตามทฤษฎีกฎหมายอาญา ผลการทดลองแสดงให้เห็นว่าแบบจำลองให้ประสิทธิภาพสูงกว่าแบบจำลองที่ใช้เทคนิคการเรียนรู้ของเครื่องเดิม เช่น Naive Bayes และ SVM เมื่อพิจารณาจากค่า F1 นอกจากนี้ แบบจำลองยังให้ประสิทธิภาพสูงในการทำนายประเด็นในคดีอาญาบางประเด็นซึ่งมีผลต่อการทำนายผลคำตัดสินในคดีอาญาด้วย นอกจากนั้น ผลการทดลองสะท้อนให้เห็นว่า การใช้โครงข่ายประตูวกกลับสองทิศทางร่วมด้วยกลไกจุดสนใจสามารถสร้างตัวแทนข้อความที่ดีกว่าแบบจำลองดั้งเดิมที่มีลักษณะเดียวกันกับแบบจำลองถุงคำ (BoW) ตลอดจนโครงข่ายประสาทเทียมแบบโมดูลสามารถจำลองโครงสร้างความรับผิดทางอาญาได้ [5]

วิธีการเรียนรู้เชิงลึกที่ใช้ในงานวิจัยนี้มีวัตถุประสงคเพื่อพัฒนาแบบจำลองการดึงดูดการเดินทางโดยใชข้อมูลเครือข่ายสังคมออนไลน์ โดยงานวิจัยนี้ใชขอมูลการเดินทางเข้าสู่พื้นที่อุทยานแห่งชาติเป็นกรณีศึกษา ทำการพัฒนาแบบจำลองโดยใชวิธีการเรียนรู้เชิงลึก และเปรียบเทียบกับการวิเคราะห์ด้วยวิธีวิเคราะห์ความถดถอยและวิธีโครงข่ายประสาทเทียม ผลการวิจัยพบว่า ข้อมูลจากการเช็คอินสามารถนำมาสรางแบบจำลองการดึงดูดการเดินทางได้ เมื่อเปรียบเทียบแบบจำลองทั้งสามแบบพบว่าแบบจำลองที่ได้จากวิธีการเรียนรู้เชิงลึกให้ความถูกต้องในการพยากรณ์สูงที่สุด โดยให้ค่าเฉลี่ยเปอร์เซ็นต์ของความคลาดเคลื่อนสัมบูรณ์เท่ากับ 43.99 ทั้งค่าที่ได้ต่ำกว่าวิธีวิเคราะห์ความถดถอยเชิงเสน และวิธีโครงข่ายประสาทเทียมซึ่งมีค่าเท่ากับ 386.48 และ 88.97 ตามลำดับ ทั้งนี้การ

พัฒนาแบบจำลองโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์จะช่วยลดค่าใช้จ่ายในการสำรวจข้อมูลลงได้อย่างมาก [25]

การประยุกต์ใช้เทคโนโลยีการเรียนรู้เชิงลึกในการจำแนกข้อมูลถนนจากภาพถ่าย Drone เพื่อการสำรวจถนนในเขตชนบท วัตถุประสงค์ของงานวิจัยครั้งนี้ได้ศึกษาเทคโนโลยีการเรียนรู้เชิงลึก (Deep Learning) ที่สามารถจำแนกและวิเคราะห์ข้อมูลทางภูมิศาสตร์ (Spatial Data) ในภาพถ่ายดาวเทียมหรือภาพถ่ายUAV เพื่อปรับปรุง (Update) ข้อมูลถนนใน OpenStreetMap (OSM) ซึ่งจะเลือกพื้นที่ที่ยังขาดข้อมูลในส่วนนี้ ไม่ว่าจะเป็นพื้นที่อยู่ไกลออกไปจากตัวเมือง เช่นตามเขตชนบท หรือในพื้นที่ ที่ข้อมูลภาพถ่ายจาก Google Map ยังไม่มีการปรับปรุง ซึ่งเทคนิคการเรียนรู้เชิงลึกนี้จะใช้ Python เพื่อจำแนกข้อมูลในภาพถ่ายและแปลงให้อยู่ในรูปแบบ Shapefile ในปัจจุบันเทคโนโลยีการเรียนรู้เชิงลึกนี้ถูกนำมาใช้ประโยชน์ในหลาย ๆ ด้าน ไม่ว่าจะเป็นพาหนะไร้คนขับอย่างเช่น Google Car ซึ่งสามารถวิ่งไปบนท้องถนน สามารถรับรู้สถานการณ์ต่าง ๆ บนท้องถนนและสามารถขับเคลื่อน และหลบหลีกสิ่งกีดขวางได้อย่างปลอดภัย หรือการ Tag เพื่อนของเราในรูปที่เราโพสต์บน Facebook ล้วนต่างใช้เทคโนโลยีการเรียนรู้เชิงลึกทั้งหมด ผู้วิจัยจึงได้มองเห็นศักยภาพของระบบเทคโนโลยีในปัจจุบันที่น่าจะนำมาใช้ในการแก้ปัญหาในส่วนนี้ และจากผลงานวิจัยนี้จึงทำให้สามารถจำแนกและวิเคราะห์ข้อมูลถนนในภาพถ่ายเพื่ออัพเดตข้อมูล OSM ได้อย่างถูกต้องรวดเร็วและแม่นยำขึ้น [26]

ความแม่นในการนำปัญญาประดิษฐ์ที่มีการเรียนรู้เชิงลึกมาทำหน้าที่แบ่งแยกพื้นที่บาดแผลเรื้องรังเป็นสิ่งที่จำเป็นสำหรับการประเมินและดูความก้าวหน้าในการฟื้นตัวของสภาพบาดแผล อย่างไรก็ตามความแม่นในการวัดขนาดบาดแผลขึ้นอยู่กับความชำนาญของผู้วัดขนาด กล่าวคือผลการวัดอาจเปลี่ยนไปมากเมื่อเปลี่ยนผู้วัดขนาด ทำให้เกิดเป็นความคลาดเคลื่อนและทำให้ประเมินความก้าวหน้าในการรักษาผิดพลาด จากปัญหาเหล่านี้จึงได้มีแนวคิดนำปัญญาประดิษฐ์ที่มีการเรียนรู้เชิงลึกสำหรับการแบ่งแยกเชิงความหมายมาแก้ไขปัญหาที่เกิดขึ้น โดยที่ปัญญาประดิษฐ์จะทำหน้าที่แบ่งแยกภาพบาดแผลเพื่อให้ได้พื้นที่บาดแผลจริงออกมา แล้วนำพื้นที่บาดแผลนั้นไปเข้าสู่กระบวนการวัดและประเมินสภาพต่อไปในอนาคต แต่น่าเสียดายที่การแบ่งส่วนเชิงความหมายในงานก่อนหน้าให้ไม่ได้ผลลัพธ์ที่น่าพอใจสำหรับงานด้านการแบ่งแยกภาพบาดแผล ถึงแม้ว่าจะมีชุดข้อมูลการฝึกที่มีขนาดใหญ่ก็ตาม เพื่อตอบสมมุติฐานที่เกิดขึ้นในการทดลองทางผู้วิจัยได้มีการเพิ่มความหลากหลายของสีด้วยการขยายข้อมูลด้วยตัวแบบรูปแบบสีจำนวน 6 รูปแบบ และนอกจากนี้ยังมีการแบ่งแยกประเภทเนื้อเยื่อบาดแผลออกเป็น 3 ประเภทได้แก่ เนื้อเยื่อแผลเนื้อแดง (granulation) ,

หนอง (slough) และเนื้อเยื่อแผลเนื้อตาย (necrosis) ทั้งในชุดข้อมูลการฝึกและชุดข้อมูลการทดสอบ จากผลการทดลองแสดงให้เห็นว่าการเพิ่มความหลากหลายของสีของภาพแผลช่วยเพิ่มความแม่นการแบ่งแยกพื้นที่พื้นที่บาดแผลได้ทำให้กล่าวได้ว่าสีส่งผลกระทบต่อการแบ่งส่วนบาดแผลอย่างมีนัยสำคัญ และในงานวิจัยนี้มีความแม่นการแบ่งแยกพื้นที่บาดแผลใกล้เคียงวิธีการแบบบก่อนหน้าถึงแม้ว่าจะมีชุดข้อมูลการฝึกขนาดเล็กก็ตาม [27]

2.2.2 งานวิจัยต่างประเทศ

วิธีการสำหรับการจำแนกข้อมูลขนาดใหญ่มากโดยใช้การวิเคราะห์กลุ่มด้วยวิธีเคมีนและวิธี Multi-kernel SVM เพื่อลดขนาดของข้อมูลฝึกและลดเวลาในการฝึก โดยการรวมเทคนิคการจัดกลุ่มของวิธีเคมีนและวิธี Multi-kernel SVM โดยเริ่มจากการลดขนาดของข้อมูลด้วยคุณสมบัติของวิธีเคมีนและการตรวจหาค่าผิดปกติ ขั้นตอนที่สองคือ ใช้วิธี Multi-kernel SVM ในการจำแนกประเภท ผลลัพธ์ที่ได้แสดงให้เห็นว่าวิธีการเลือกข้อมูลฝึกที่นำเสนอสามารถลดขนาดของข้อมูลฝึก อีกทั้งยังลดเวลาที่ใช้ในการฝึก และสามารถรักษาประสิทธิภาพความแม่นได้ดี [3]

จากการศึกษางานวิจัยพบว่า งานวิจัยเกี่ยวกับการจำแนกข้อมูลขนาดใหญ่ที่ผ่านมา [4, 27] โดยงานวิจัยส่วนมากศึกษาการจำแนกข้อมูลขนาดใหญ่ด้วยวิธีการซัพพอร์ตเวกเตอร์แมชชีน (SVM) และวิธีโครงข่ายประสาทเทียม (ANN) ซึ่งมักมีปัญหาในด้านของขนาดข้อมูลฝึกที่ใหญ่มากหรือจำนวนเยอะมากจนเกินไปและส่งผลต่อเวลาในการประมวลผล ดังนั้นผู้วิจัยจึงมีจุดประสงค์ที่จะศึกษาการจำแนกข้อมูลขนาดใหญ่มาก เพื่อลดปัญหาสำหรับการใช้ข้อมูลฝึกจำนวนมาก โดยจะทำการลดขนาดข้อมูลฝึกด้วยการรวมเทคนิคการจัดกลุ่มของวิธีเคมีนและวิธีการเรียนรู้เชิงลึก โดยเริ่มจากการลดขนาดข้อมูลด้วยคุณสมบัติของวิธีเคมีนและการตรวจหาค่าผิดปกติ ในขั้นตอนต่อมาจะใช้วิธีการเรียนรู้เชิงลึกในการจำแนกประเภทเพื่อให้ได้ข้อมูลฝึกที่มีขนาดลดลง แต่ยังคงมีประสิทธิภาพและมีความแม่นสูง จากนั้นทำการเปรียบเทียบประสิทธิภาพความแม่นระหว่างวิธีการที่นำเสนอกับวิธีการเรียนรู้เชิงลึกที่ใช้การสุ่มข้อมูลฝึกขนาดเท่ากับวิธีการที่นำเสนอ และเปรียบเทียบประสิทธิภาพความแม่นระหว่างวิธีการที่นำเสนอกับวิธีการเรียนรู้เชิงลึกที่ใช้ข้อมูลฝึกขนาด 80% กับ 90% ของจำนวนข้อมูลทั้งหมด

บทที่ 3
วิธีการดำเนินการวิจัย

การวิจัยเรื่องการจำแนกข้อมูลขนาดใหญ่มากโดยใช้การจัดกลุ่มด้วยวิธีเคมีนและวิธีการเรียนรู้เชิงลึก มีวัตถุประสงค์เพื่อทำให้ข้อมูลฝึกมีขนาดลดลง แต่ยังคงมีประสิทธิภาพและมีความแม่นสูง โดยกำหนดขนาดของข้อมูลที่มีขนาดใหญ่มาก นั่นคือจำนวนข้อมูล (N)×คุณลักษณะ (Feature) ≥ 500,000 และกำหนดวิธีการที่ใช้ในการศึกษาดังนี้

3.1 วิธีการจัดกลุ่มด้วยคุณสมบัติของวิธีเคมีนและการตรวจหาค่าผิดปกติ

3.2 วิธีการเรียนรู้เชิงลึก

3.1 วิธีการจัดกลุ่มด้วยคุณสมบัติของวิธีเคมีนและการตรวจหาค่าผิดปกติ

3.1.1 การเก็บรวบรวมข้อมูล

ชุดข้อมูลขนาดใหญ่มากที่ได้จากการสร้างข้อมูล (Generate Data)

ข้อมูลที่ใช้ในการศึกษาครั้งนี้ (Feature) มี 3 การแจกแจง โดยมีค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สันระหว่าง Feature ไม่เกิน ± 0.5 ได้แก่

1) การแจกแจงปรกติมาตรฐาน (Standard Normal Distribution: N(0,1))

2) การแจกแจงแบบเลขชี้กำลัง (Exponential Distribution: exp(1))

3) การแจกแจงเอกรูป (Uniform Distribution: U(0,1))

ในแต่ละการแจกแจงทำการสร้างข้อมูล (N × Feature) 4 ขนาด ประกอบไปด้วย

1) ขนาด 1,000,000×4

2) ขนาด100,000×9

3) ขนาด 30,000×29

4) ขนาด 7,000×74

การสร้าง Target ของแต่ละการแจกแจง ทั้ง 4 ขนาด มีรายละเอียดดังนี้

1) หาค่าเฉลี่ยของ Feature ในทุก Case

2) นำค่าเฉลี่ยที่ได้จาก ข้อ 1) จัดให้อยู่ในรูปของ Quantile

3) นำ Quantile ที่ได้จาก ข้อ 2) มาทำการแบ่ง Class (ในการวิจัยนี้กำหนดเป็น 3 Class แต่ละ Class กำหนดเป็น 45% 10% และ 45% ตามลำดับ)

ชุดข้อมูลจริงที่มีขนาดใหญ่มากประกอบไปด้วย 2 ชุดข้อมูลดังนี้

1) ชุดข้อมูล Skin Segmentation โดยมีขนาดข้อมูล 245,057×3 และมี Class จำนวน 2 Class คิดเป็นสัดส่วน 0.21 : 0.79 ตามลำดับ จากฐานเก็บข้อมูล UCI สืบค้นจาก https://archive.ics.uci.edu

2) ชุดข้อมูล Coil 2000 โดยมีขนาดข้อมูล 9,822×84 และมี Class จำนวน 2 Class คิดเป็นสัดส่วน 0.94 : 0.06 ตามลำดับ จากฐานเก็บข้อมูล KEEL สืบค้นจาก https://sci2s.ugr.es/keel

3.1.2 ขั้นตอนการวิเคราะห์กลุ่มด้วยคุณสมบัติของวิธีเคมีนและการตรวจหาค่าผิดปกติ

ขั้นตอนการวิเคราะห์กลุ่มด้วยคุณสมบัติของวิธีเคมีนและการตรวจหาค่าผิดปกติ

1) ตรวจสอบความสมบูรณ์ของข้อมูล เช่น กรณีข้อมูลมี Missing Data หรือมี NA ให้ทำการตัดข้อมูล Case นั้นออก

2) กำหนดสัดส่วน (Proportion) ในการสุ่มข้อมูลเพื่อเป็นข้อมูลฝึก

สำหรับข้อมูลที่มีขนาดใหญ่มาก หากนำข้อมูลทั้งหมดมาใช้เพื่อเป็นข้อมูลฝึกจะทำให้เวลาในการประมวลผลนานมาก เมื่อข้อมูลมีขนาดใหญ่มากสามารถกำหนดสัดส่วนในการสุ่มข้อมูลเพียงเล็กน้อยเพื่อเป็นข้อมูลฝึก เช่น สุ่มข้อมูลฝึกด้วยสัดส่วน 0.05 ถึง 0.10 นั่นคือ 5% ถึง 10% ของข้อมูลทั้งหมด หรืออาจกำหนดสัดส่วนในการสุ่มข้อมูลฝึกมากกว่า 10% เพื่อประสิทธิภาพความแม่นอาจสูงขึ้น ทั้งนี้ส่งผลให้เวลาในการประมวลผลนานมากขึ้นเช่นกัน

3) ทำการวิเคราะห์กลุ่มด้วยวิธีเคมีนและลดขนาดข้อมูล

- กำหนดจำนวนกลุ่ม (Km) และจำนวนรอบในการทำซ้ำ (RT)
- หาระยะห่างระหว่างจุดศูนย์กลางของแต่ละกลุ่ม
- เก็บ Case ที่ใกล้สุดและไกลสุดจากจุดศูนย์กลางของแต่ละกลุ่มในทุก K และ RT
- ลดขนาดของข้อมูลโดยการลบข้อมูลที่ซ้ำออก

4) คำนวณค่าผิดปกติของชุดข้อมูลโดยสูตรดังนี้ [3]

$$KSE(p_j) = \frac{1}{n-1}\sum_{i=1}^{n} KS(p_j - p_i) \qquad (3.1)$$

โดยที่ $$KS(p_j - p_i) = sup_x \left| Fp_j(x) - Fp_i(x) \right|$$

เมื่อ Fp_j คือ ระยะทางจากจุด j ไปยังจุดอื่น ๆ ที่รวบรวม

Fp_i คือ ระยะทางจากจุด i ไปยังจุดอื่น ๆ ที่รวบรวม

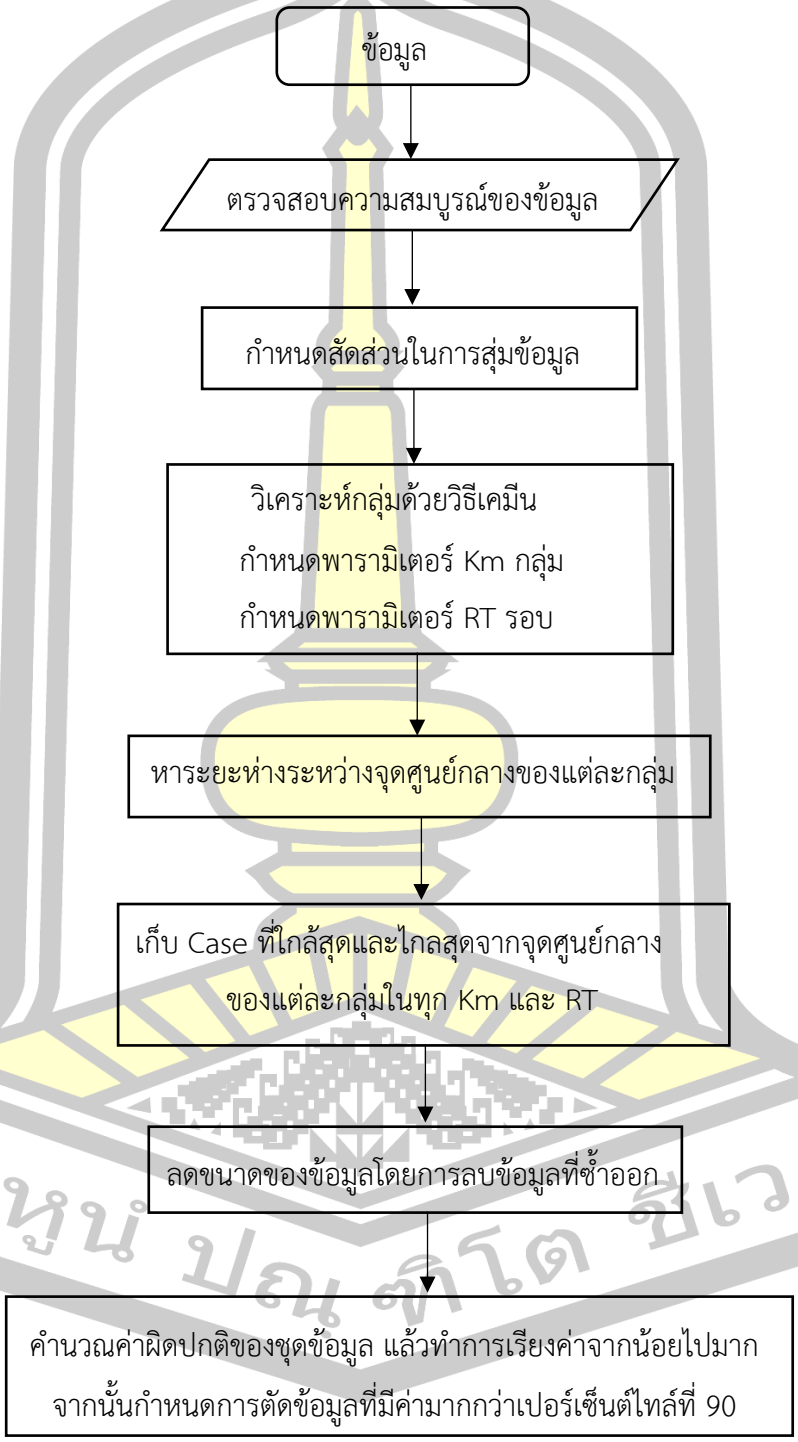
จากการคำนวณ $KSE(p_j)$ ทำการเรียงค่าจากน้อยไปมาก จากนั้นทำการตัดข้อมูลที่มีค่ามากกว่าตำแหน่งเปอร์เซ็นต์ไทล์ต่าง ๆ เช่น ตัดข้อมูลที่มีค่ามากกว่าตำแหน่งเปอร์เซ็นต์ไทล์ที่ 90 เป็นต้น เพื่อกำจัดค่าสูงสุดและลดขนาดของข้อมูลฝึกในการคำนวณหาตำแหน่งเปอร์เซ็นต์ไทล์โดยสูตรดังนี้ [28]

$$P_r = \frac{r}{100}(n+1) \tag{3.2}$$

เมื่อ P_r คือ ตำแหน่งของข้อมูล ณ เปอร์เซ็นต์ไทล์ที่ r

n คือ จำนวนข้อมูลทั้งหมด

สามารถสรุปแผนผังการวิเคราะห์ข้อมูลด้วยคุณสมบัติของวิธีเคมีนและการตรวจหาค่าผิดปกติได้ดังนี้



ภาพที่ 15 แผนผังการวิเคราะห์ข้อมูลด้วยคุณสมบัติของวิธีเคมีนและการตรวจหาค่าผิดปกติ

3.2 วิธีการเรียนรู้เชิงลึก

3.2.1 การวิเคราะห์ข้อมูลด้วยวิธีการเรียนรู้เชิงลึก

วิธีการเรียนรู้เชิงลึกจะทำการคำนวณผ่าน ANN โดย ANN สามารถคำนวณได้ดังนี้

$$y_j = f\left(\sum_{i=1}^{n} x_i w_{ij} + \theta_j\right) \tag{3.3}$$

เมื่อ y_j คือ ผลลัพธ์ในชั้นซ่อน หรือข้อมูลส่งออกในชั้นซ่อนโหนดที่ j

x_i คือ ข้อมูลนำเข้าโหนดที่ i ในชั้นอินพุต

w_{ij} คือ น้ำหนักบนเส้นเชื่อมระหว่างโหนดที่ i ในชั้นอินพุตและโหนดที่ j ในชั้นซ่อน

θ_j คือ ค่า Bias ของโหนดที่ j ในชั้นซ่อน

n คือ จำนวนโหนดทั้งหมดของชั้นอินพุต

3.2.2 ขั้นตอนการวิเคราะห์ข้อมูลด้วยวิธีการเรียนรู้เชิงลึก

1) วิเคราะห์ข้อมูลด้วยวิธีการเรียนรู้เชิงลึกโดยใช้ข้อมูลฝึกที่ผ่านคุณสมบัติของวิธีเคมีนและการตรวจหาค่าผิดปกติ (วิธีการที่นำเสนอ) โดยกำหนดชั้นซ่อน (Hidden Layer) เป็น 3 ชั้นซ่อน โดยในแต่ละชั้นซ่อนกำหนดโหนดเป็น 5 10 และ 20 ซึ่งรวมทั้งหมดเป็น 27 กรณี ดังตาราง 1

ตาราง 1 แสดงการกำหนดชั้นซ่อนและโหนด

ชั้นซ่อน	โหนด
1, 2, 3	(5, 5, 5), (5, 5, 10), (5, 5, 20), (5, 10, 5), (5, 10, 10), (5, 10, 20), (5, 20, 5), (5, 20, 10), (5, 20, 20), (10, 5, 5), (10, 5, 10), (10, 5, 20), (10,10,5), (10,10,10), (10,10,20), (10,20,5), (10,20,10), (10,20,20), (20,5,5), (20,5,10), (20,5,20), (20,10,5), (20,10,10), (20,10,20), (20,20,5), (20,20,10), (20,20,20)

2) หาประสิทธิภาพความแม่นของวิธีการที่นำเสนอในข้อ 1 โดยทำการเปรียบเทียบประสิทธิภาพความแม่นจากวิธีการเรียนรู้เชิงลึกที่ใช้การสุ่มข้อมูลฝึกขนาดเท่ากับวิธีการที่นำเสนอ และเปรียบเทียบกับวิธีการเรียนรู้เชิงลึกที่ใช้ข้อมูลฝึกขนาด 80% กับ 90% ของจำนวนข้อมูลทั้งหมด

เกณฑ์ในการวัดประสิทธิภาพความแม่นยำ

ในการวัดประสิทธิภาพความแม่นในการศึกษาครั้งนี้ใช้เกณฑ์การวัดด้วยค่าความแม่น (Accuracy) และเกณฑ์ในการทำนาย Area Under Curve (AUC) โดยจะใช้ค่าเฉลี่ยของ Accuracy และ AUC จากวิธีการเรียนรู้เชิงลึกทั้ง 27 กรณี มีรายละเอียดดังนี้

1) เกณฑ์การวัดด้วยค่าความแม่น

ในการทดสอบประสิทธิภาพความแม่นยำ โดยใช้เกณฑ์ในการวัดด้วยค่าความแม่น โดยคำนวณจากค่าในแนวเส้นทแยงมุมของเมทริกซ์ความสับสน (Confusion Matrix: CM) ได้ดังตาราง 2

ตาราง 2 เมทริกซ์ความสับสน แบบ 2×2

	Predicted Positive	**Predicted Negative**
Positive Class	True Positive (TP)	False Negative (FN)
Negative Class	False Positive (FP)	True Negative (TN)

สามารถคำนวณได้โดยสูตร ดังนี้ [30]

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{3.4}$$

โดยที่ $Accuracy$ คือ ค่าอยู่ระหว่าง 0 - 1 เมื่อค่าเข้าใกล้ 1 นั่นคือตัวแบบสามารถจำแนกประเภทได้ดีมาก

TP คือ จำนวนที่จำแนกประเภทให้อยู่ใน Positive Class เมื่อข้อมูลจริงอยู่ใน Positive Class

TN คือ จำนวนที่จำแนกประเภทให้อยู่ใน Negative Class เมื่อข้อมูลจริงอยู่ใน Negative Class

FP คือ จำนวนที่จำแนกประเภทให้อยู่ใน Positive Class เมื่อข้อมูลจริงอยู่ใน Negative Class

FN คือ จำนวนที่จำแนกประเภทให้อยู่ใน Negative Class เมื่อข้อมูลจริงอยู่ใน Positive Class

ตาราง 3 เมทริกซ์ความสับสน แบบ 3×3

Confusion Matrix		Predicted		
		Class 1	Class 2	Class 3
Actual	Class 1	A	B	C
	Class 2	D	E	F
	Class 3	G	H	I

สามารถคำนวณได้โดยสูตร ดังนี้ [31]

จาก (3.4) $$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} = \frac{A + (E + I)}{A + (B + C) + (E + I) + (D + G)} \quad (3.5)$$

2) เกณฑ์ในการทำนาย Area Under Curve (AUC)

ในการวัดผลด้วยค่า AUC นิยมใช้วัดผลตัวแบบในทุกงานโดย AUC มีค่าอยู่ระหว่าง 0 - 1 เมื่อค่าเข้าใกล้ 1 นั่นหมายความว่าตัวแบบในภาพรวมสามารถจำแนกประเภทได้ดีมาก ซึ่งสามารถคำนวณได้โดยสูตรดังนี้ [32]

$$AUC = \frac{Sensitivity + Specificity}{2} \quad (3.6)$$

โดยที่ $Sensitivity$ คือ $\frac{TP}{TP + FN}$

$Specificity$ คือ $\frac{TN}{TN + FP}$

กรณี Confusion Matrix แบบ 3×3 สามารถคำนวณได้โดยสูตรดังนี้ [31]

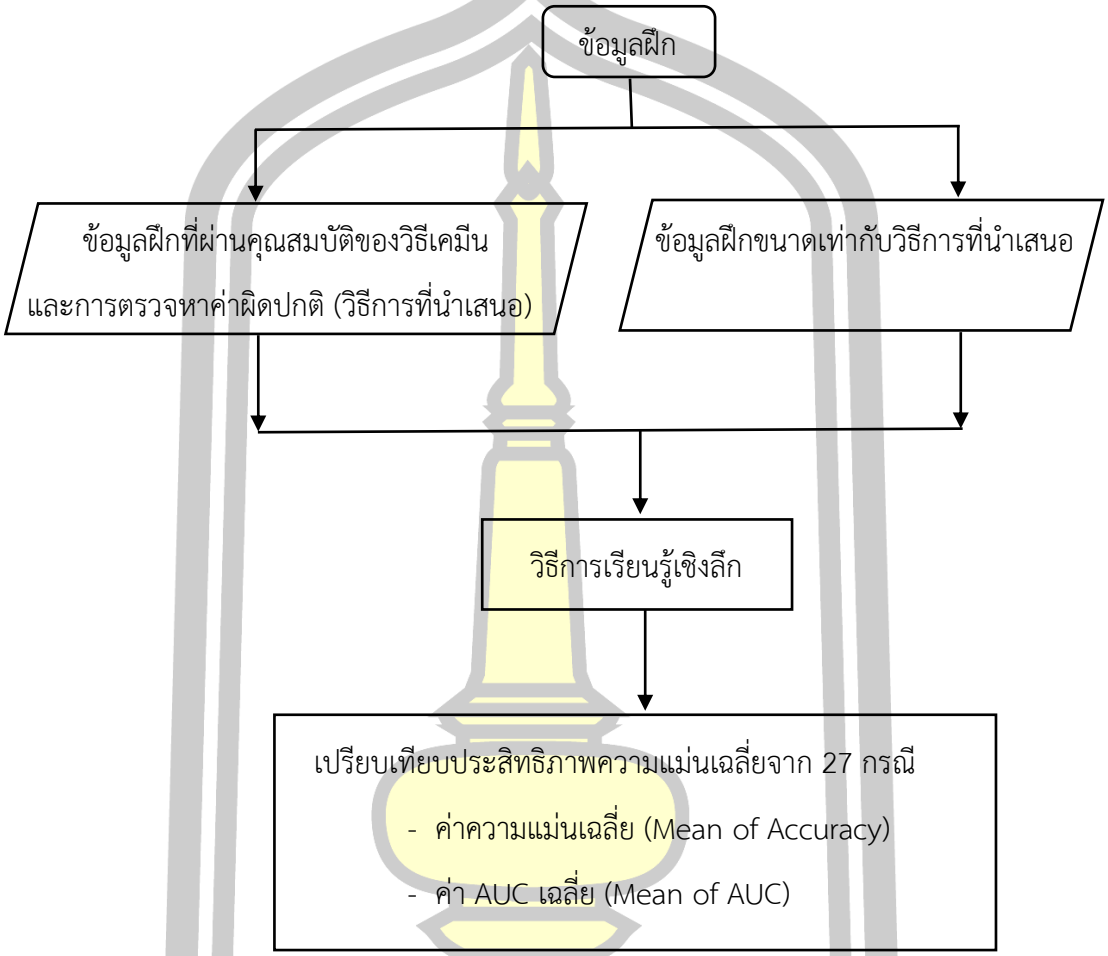
จากตาราง 3 $Sensitivity$ คือ $\frac{A}{A+B+C}$

$Specificity$ คือ $\frac{E+I}{D+G+E+I}$

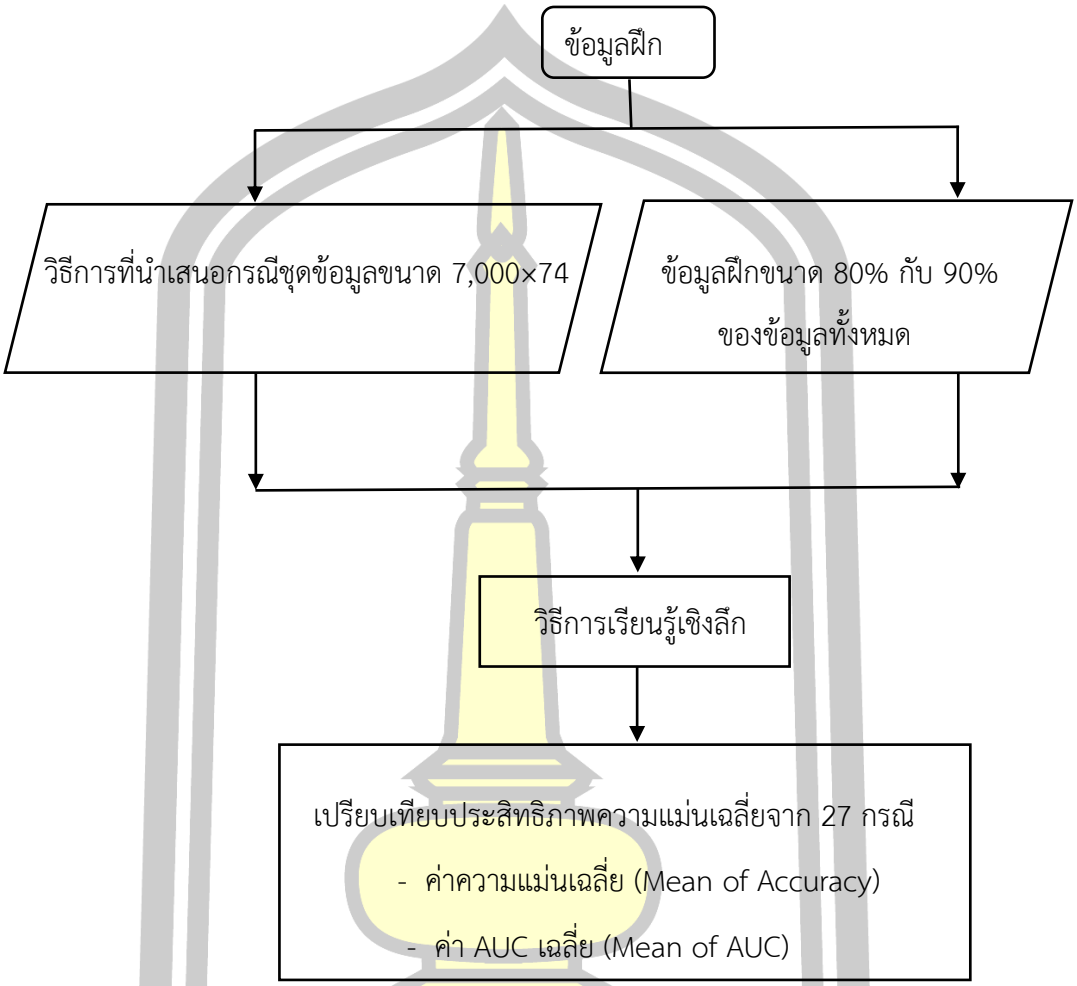
สามารถสรุปได้ดังเกณฑ์ต่อไปนี้ [33]

- $0.50 \leq AUC < 0.70$ คือ ตัวแบบมีประสิทธิภาพต่ำ
- $0.70 \leq AUC < 0.80$ คือ เกณฑ์มาตรฐานสำหรับตัวแบบส่วนใหญ่
- $0.80 \leq AUC < 0.90$ คือ ตัวแบบทำงานได้ดี
- $AUC > 0.90$ คือ ตัวแบบทำงานได้ดีมาก

สามารถสรุปแผนผังการวิเคราะห์ข้อมูลด้วยวิธีการเรียนรู้เชิงลึกได้ดังภาพที่ 16 และภาพที่ 17



ภาพที่ 16 แผนผังการวิเคราะห์ข้อมูลด้วยวิธีการเรียนรู้เชิงลึกที่ใช้ข้อมูลฝึกขนาดเท่ากับวิธีการที่นำเสนอ



ภาพที่ 17 แผนผังการวิเคราะห์ข้อมูลด้วยวิธีการเรียนรู้เชิงลึกที่ใช้ข้อมูลฝึกขนาด 80% กับ 90% เทียบกับวิธีการที่นำเสนอ

ในการเปรียบเทียบการใช้ข้อมูลฝึกที่ผ่านคุณสมบัติของวิธีเคมีนและการตรวจหาค่าผิดปกติกับข้อมูลฝึกที่ทำการสุ่มขนาดต่าง ๆ ด้วยวิธี K-fold Cross Validation โดยกำหนด K=10 หมายถึงการแบ่งข้อมูลออกเป็น 10 ส่วน ในแต่ละส่วนมีจำนวนข้อมูลเท่า ๆ กัน โดยสร้างแบบจำลองจากข้อมูล 9 ชุด อีก 1 ชุดที่เหลือใช้สำหรับทดสอบ วนซ้ำการทำงานนี้จนกระทั่งข้อมูลทุกส่วนถูกใช้เป็นชุดทดสอบ [34]

บทที่ 4
ผลการวิจัย

ในบทนี้จะนำเสนอผลการวิจัยของวิธีการเรียนรู้เชิงลึกโดยใช้ข้อมูลฝึกที่ผ่านคุณสมบัติของวิธีเคมีนและการตรวจหาค่าผิดปกติ จะนำเสนอเป็น 2 ส่วน โดยมีรายละเอียดดังนี้

4.1 การเปรียบเทียบประสิทธิภาพของวิธีการที่นำเสนอกับวิธีการเรียนรู้เชิงลึกที่ใช้การสุ่มข้อมูลฝึกขนาดเท่ากับวิธีการที่นำเสนอทั้งจากข้อมูลที่สร้างขึ้นและข้อมูลจริง

4.2 เปรียบเทียบประสิทธิภาพของการจำแนกประเภทและเวลาของวิธีการที่นำเสนอกับวิธีการเรียนรู้เชิงลึกโดยใช้ข้อมูลฝึกขนาด 80% กับ 90%

4.1 การเปรียบเทียบประสิทธิภาพของวิธีการที่นำเสนอกับวิธีการเรียนรู้เชิงลึกที่ใช้การสุ่มข้อมูลฝึกขนาดเท่ากับวิธีการที่นำเสนอทั้งจากข้อมูลที่สร้างขึ้นและข้อมูลจริง

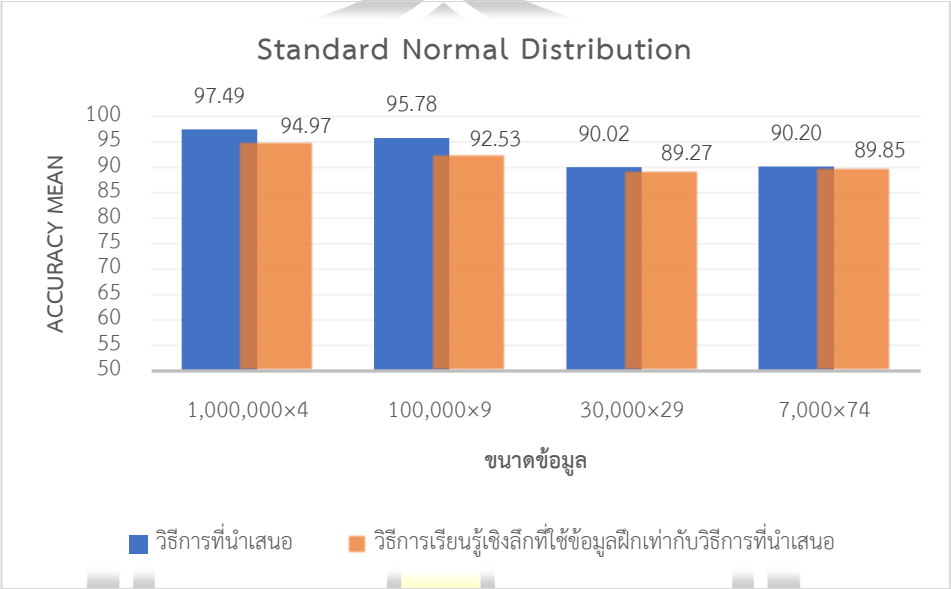
ในส่วนนี้จะแสดงประสิทธิภาพความแม่นของวิธีการที่นำเสนอกับวิธีการเรียนรู้เชิงลึกที่ใช้การสุ่มข้อมูลฝึกขนาดเท่ากับวิธีการที่นำเสนอ ค่าใน [.] จะแสดงให้เห็นถึงขนาดของข้อมูลฝึกที่ลดลงแต่ยังคงประสิทธิภาพความแม่นสูง โดยกำหนดการตัดข้อมูลที่มีค่ามากกว่าตำแหน่งเปอร์เซ็นต์ไทล์ที่ 90 ใน 1 รอบของการทำซ้ำ ดังตาราง 4 และตาราง 5 อีกทั้งยังแสดงชั้นซ่อนที่กำหนดพร้อมทั้งค่า Confusion Matrix รวมถึงค่าความแม่นและค่า AUC เฉลี่ยจาก 27 กรณี ใน 1 รอบของการทำซ้ำ ดังตาราง 6

ตาราง 4 แสดงประสิทธิภาพของวิธีการที่นำเสนอจากการสร้างข้อมูล โดยแสดงถึงขนาดข้อมูลฝึกและประสิทธิภาพความแม่นเมื่อเทียบกับวิธีการเรียนรู้เชิงลึกที่ใช้การสุ่มข้อมูลฝึกขนาดเท่ากับวิธีการที่นำเสนอ

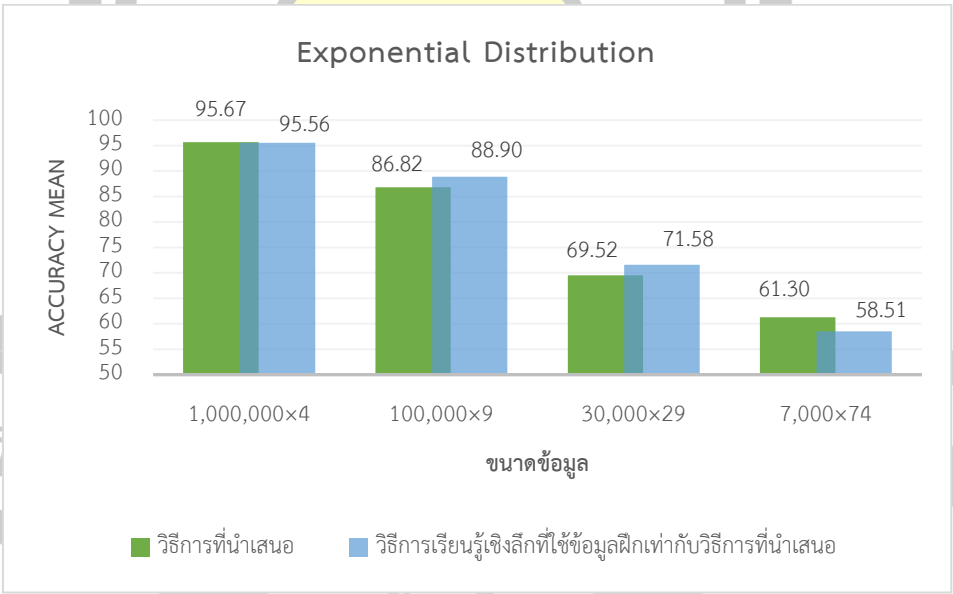
การแจกแจงของข้อมูล	**ขนาดข้อมูล (N×Feature)**	**พารามิเตอร์ สัดส่วน**	**ขนาดข้อมูลฝึก Case (%)**	**ประสิทธิภาพความแม่น**	
				Mean of Accuracy %	**Mean of AUC**
N(0,1)	1,000,000×4	0.05	104 (0.01%)	97.4878 [94.9744]	0.9735 [0.9659]
	100,000×9	0.10	90 (0.09%)	95.7757 [92.5282]	0.92284 [0.9075]
	30,000×29	0.10	50 (0.17%)	90.0225 [89.2700]	0.8280 [0.8333]
	7,000×74	0.10	48 (0.69%)	90.1968 [89.8467]	0.8026 [0.8351]
exp(1)	1,000,000×4	0.05	109 (0.01%)	95.6735 [95.5576]	0.9259 [0.9568]
	100,000×9	0.10	93 (0.09%)	86.8219 [88.9000]	0.8231 [0.8722]
	30,000×29	0.10	79 (0.26%)	69.5191 [71.5839]	0.6938 [0.6775]
	7,000×74	0.10	61 (0.87%)	61.3046 [58.5077]	0.5946 [0.5986]
U(0,1)	1,000,000×4	0.05	55 (0.01%)	**95.7708 [89.6742]**	**0.9434 [0.9184]**
	100,000×9	0.10	99 (0.10%)	90.4185 [93.0459]	0.8663 [0.9098]
	30,000×29	0.10	78 (0.26%)	81.4312 [80.9128]	0.7843 [0.7817]
	7,000×74	0.10	60 (0.86%)	75.2455 [72.9768]	0.7069 [0.7151]

หมายเหตุ : กำหนดพารามิเตอร์ Km และ RT = 10 ในการหาข้อมูลฝึก; ค่าใน [·] คือวิธีการเรียนรู้เชิงลึก

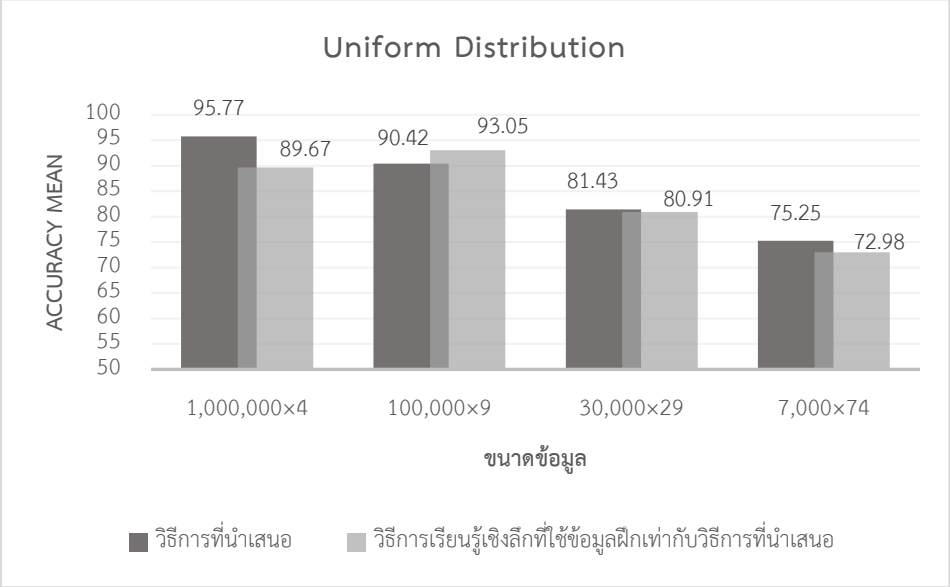
ในส่วนนี้จะแสดงค่าความแม่นเฉลี่ยจากตาราง 4 ของวิธีการที่นำเสนอและวิธีการเรียนรู้เชิงลึกที่ใช้ข้อมูลฝึกขนาดเท่ากับวิธีการที่นำเสนอของทั้ง 3 การแจกแจง ดังภาพที่ 18 - 20



ภาพที่ 18 แสดงค่าความแม่นเฉลี่ยของชุดข้อมูลที่มีการแจกแจงปรกติมาตรฐาน

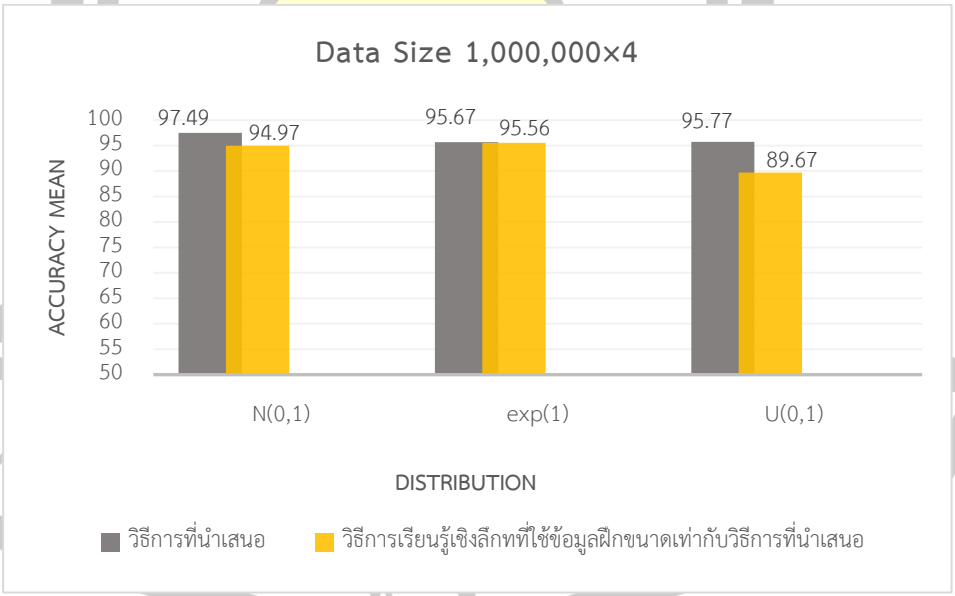


ภาพที่ 19 แสดงค่าความแม่นเฉลี่ยของชุดข้อมูลที่มีการแจกแจงแบบเลขชี้กำลัง

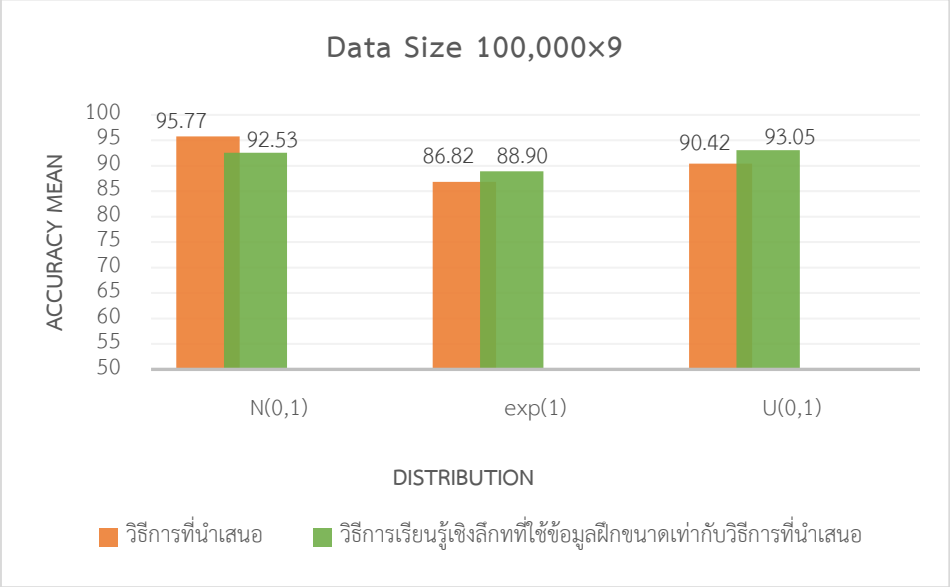


ภาพที่ 20 แสดงค่าความแม่นเฉลี่ยของชุดข้อมูลที่มีการแจกแจงเอกรูป

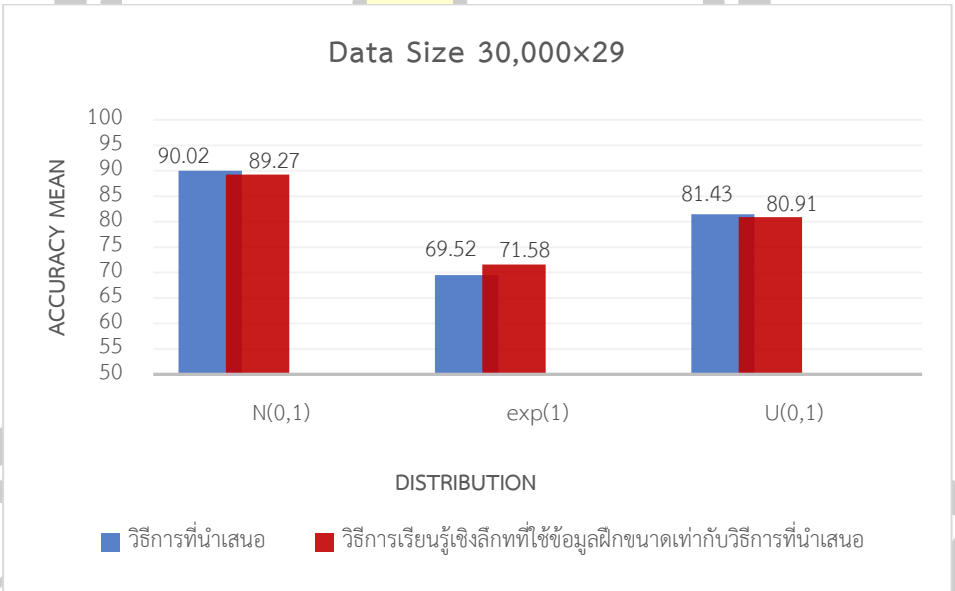
แสดงค่าความแม่นเฉลี่ยจากตาราง 4 โดยแบ่งตามขนาดข้อมูลของวิธีการที่นำเสนอและวิธีการเรียนรู้เชิงลึกที่ใช้ข้อมูลฝึกขนาดเท่ากับวิธีการที่นำเสนอของขนาดข้อมูลทั้ง 4 ขนาด ดังภาพที่ 21 - 24



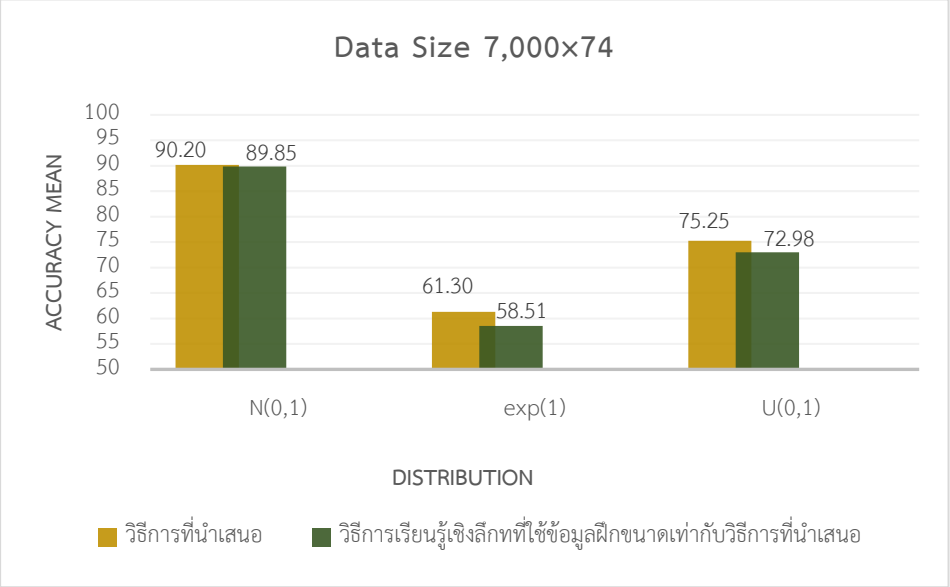
ภาพที่ 21 แสดงค่าความแม่นเฉลี่ยของทุกการแจกแจงที่ศึกษา



ภาพที่ 22 แสดงค่าความแม่นเฉลี่ยของทุกการแจกแจงที่ศึกษา



ภาพที่ 23 แสดงค่าความแม่นเฉลี่ยของทุกการแจกแจงที่ศึกษา



ภาพที่ 24 แสดงค่าความแม่นเฉลี่ยของทุกการแจกแจงที่ศึกษาใน 300 รอบของการทำซ้ำ

จากตาราง 4 แสดงค่าความแม่นเฉลี่ยที่ได้จากผลการวิเคราะห์ด้วยโหนดต่างกัน 27 กรณี โดยทำการเปรียบเทียบประสิทธิภาพของวิธีการที่นำเสนอกับวิธีการเรียนรู้เชิงลึกที่ใช้การสุ่มข้อมูลฝึกขนาดเท่ากับวิธีการที่นำเสนอ โดยข้อมูลที่ใช้ในการศึกษาเป็นข้อมูลขนาดใหญ่มากจากการสร้างข้อมูลจำนวน 3 ชุด (N × Feature) ประกอบไปด้วยชุดข้อมูลที่ Feature มีการแจกแจงปรกติมาตรฐาน Feature มีการแจกแจงแบบเลขชี้กำลัง และชุดข้อมูลที่ Feature มีการแจกแจงเอกรูป แต่ละชุดข้อมูลมี 4 ขนาด คือ1) 1,000,000×4 ขนาดที่ 2) 100,000×9 ขนาดที่ 3) 30,000×29 และขนาดที่ 4) ขนาด 7,000×74 ในแต่ละขนาดจะแสดงการกำหนดพารามิเตอร์สัดส่วนในการสุ่มข้อมูลเพื่อใช้เป็นข้อมูลฝึก ในการศึกษานี้เพื่อลดเวลาในการประมวลผลจึงกำหนดสัดส่วนในการสุ่มตั้งแต่ 0.05 ไปจนถึง 0.40 หากข้อมูลมีขนาดใหญ่มากจนเกินไป (1,000,000×4) อาจใช้ส่วนส่วนในการสุ่มมาเพียงเล็กน้อย เช่น 0.05 หรือ 5% อีกทั้งยังแสดงการกำหนดกลุ่ม (Km = 10) รวมถึงการกำหนดรอบในการทำซ้ำ (RT = 10) และกำหนดการตัดข้อมูลที่มีค่ามากกว่าตำแหน่งเปอร์เซ็นต์ไทล์ที่ 90 จากคอลัมน์ที่ 4 จะเห็นได้ว่าวิธีการที่นำเสนอสามารถลดขนาดของข้อมูลฝึกได้อย่างมากในทุกการแจกแจงที่ศึกษา โดยใช้ขนาดข้อมูลฝึกน้อยกว่า 1% ของจำนวนข้อมูลทั้งหมด

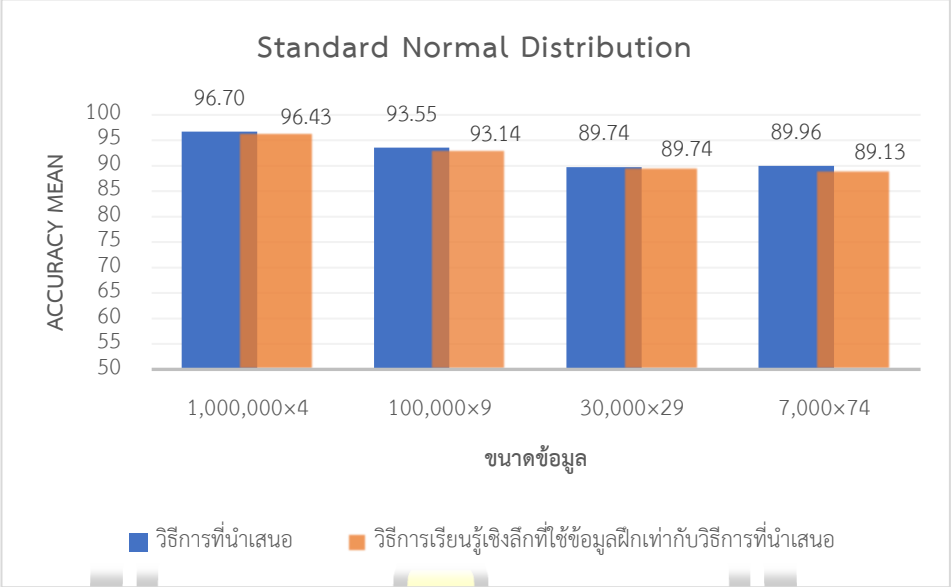
ในกรณีที่ชุดข้อมูลมีขนาด N จำนวนมาก และจำนวน Feature น้อย เช่น กรณีที่ขนาดข้อมูล 1,000,000×4 ให้ค่าความแม่นเฉลี่ยสูงมาก (> 95%) และให้ค่า AUC สูงมาก (> 0.90) ในทุกการแจกแจง โดยเฉพาะอย่างยิ่งกรณีชุดข้อมูลที่ Feature มีการแจกแจงปรกติมาตรฐาน ให้ค่า

ความแม่นเฉลี่ยสูงถึง 97.4878% อีกทั้งยังให้ค่า AUC สูงถึง 0.9735 ในขณะที่วิธีการเรียนรู้เชิงลึกที่ใช้การสุ่มข้อมูลฝึกขนาดเท่ากับวิธีการที่นำเสนอมีประสิทธิภาพความแม่นเฉลี่ย 94.9744% และค่า AUC คือ 0.9659 กรณีของชุดข้อมูลที่ Feature มีการแจกแจงแบบเลขชี้กำลังประสิทธิภาพความแม่นของวิธีการที่นำเสนอนั้นเทียบเท่ากับวิธีการเรียนรู้เชิงลึกแบบเดิมที่นำมาเปรียบเทียบ และในกรณีของชุดข้อมูลที่ Feature มีการแจกแจงเอกรูปประสิทธิภาพความแม่นของวิธีการที่นำเสนอยังคงสูงถึง 95.7708% และให้ค่า AUC สูงถึง 0.9659 แต่ในขณะที่วิธีการเรียนรู้เชิงลึกที่ใช้การสุ่มข้อมูลฝึกขนาดเท่ากับวิธีการที่นำเสนอ ซึ่งมีประสิทธิภาพความแม่นเฉลี่ย 89.6742% และค่า AUC คือ 0.9184

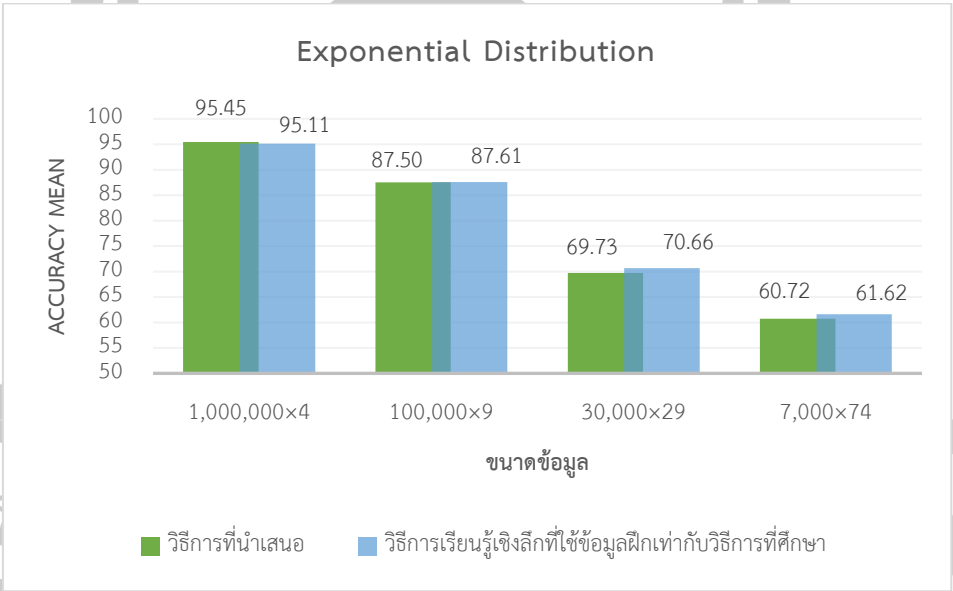
กรณีที่จำนวน Feature มากขึ้น (7,000×74) ค่าความแม่นเฉลี่ยและค่า AUC ของการจำแนกประเภทจะน้อยลงแต่ยังคงมีค่าสูงกว่าผลจากวิธีการเรียนรู้เชิงลึกที่ใช้การสุ่มข้อมูลฝึกขนาดเท่ากับวิธีการที่นำเสนอ หากพิจารณาตามการแจกแจงพบว่า ประสิทธิภาพความแม่นของวิธีการที่นำเสนอนั้นจะสูงเมื่อข้อมูล (Feature) มีการแจกแจงปรกติมาตรฐานและการแจกแจงเอกรูป และจะน้อยลงเมื่อข้อมูลมีการแจกแจงแบบเลขชี้กำลัง แต่ทั้งนี้ยังให้ค่าความแม่นเทียบเท่ากับวิธีการเรียนรู้เชิงลึกแบบเดิมที่นำมาเปรียบเทียบ

จากผลการศึกษาในตาราง 4 จะเห็นได้ว่าวิธีการที่นำเสนอให้ค่าความแม่นเฉลี่ยในการจำแนกประเภทสูงกว่าวิธีการเรียนรู้เชิงลึกที่ใช้ข้อมูลฝึกขนาดเท่ากับวิธีการที่นำเสนอในทุกกรณีของแต่ละการแจกแจง เมื่อการแจกแจงนั้นมีขนาด N จำนวนมาก และจำนวน Feature น้อย (1,000,000×4)

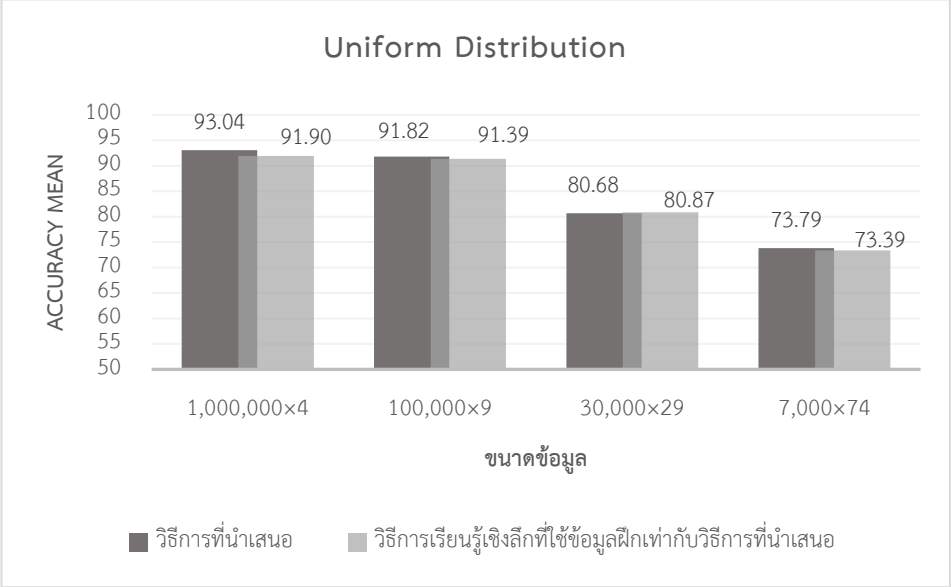
ในส่วนนี้จะเป็นการแสดงค่าความแม่นเฉลี่ยใน 300 รอบของการทำซ้ำด้วยวิธีการที่นำเสนอ โดยปกติควรทำซ้ำอยู่ที่ 10,000 รอบ แต่ในงานวิจัยนี้กำหนดการทำซ้ำที่ 300 รอบ เนื่องจากข้อจำกัดด้านเวลาที่ใช้ในการประมวลผล ทั้งนี้จะแสดงค่าความแม่นเฉลี่ยใน 300 รอบของการทำซ้ำโดยแบ่งตามการแจกแจงของวิธีการที่นำเสนอและวิธีการเรียนรู้เชิงลึกที่ใช้ข้อมูลฝึกขนาดเท่ากับวิธีการที่นำเสนอของทั้ง 3 การแจกแจง ดังภาพที่ 25 - 27



ภาพที่ 25 แสดงค่าความแม่นเฉลี่ยของชุดข้อมูลที่มีการแจกแจงปรกติมาตรฐานใน 300 รอบของการทำซ้ำ

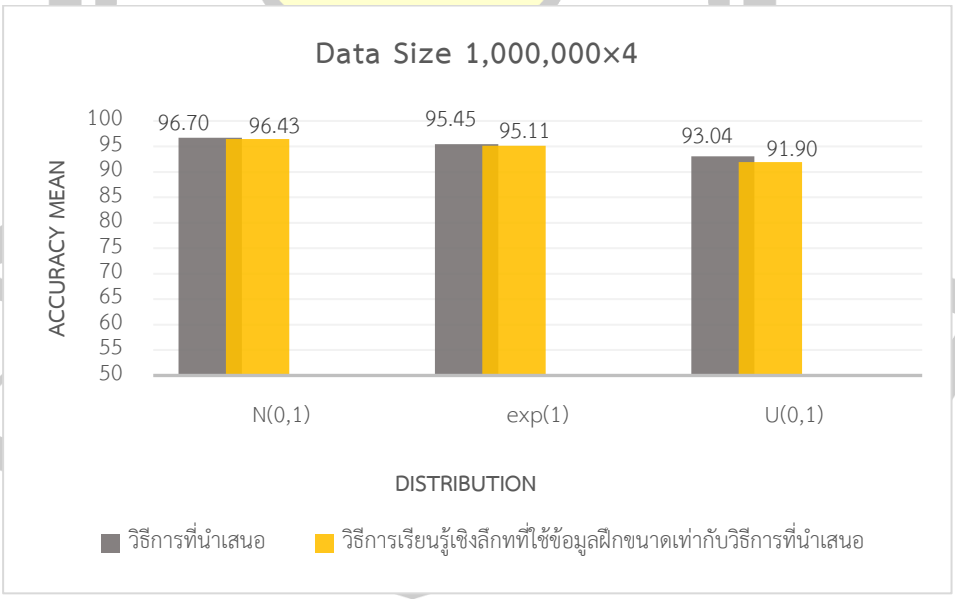


ภาพที่ 26 แสดงค่าความแม่นเฉลี่ยของชุดข้อมูลที่มีการแจกแจงแบบเลขชี้กำลังใน 300 รอบของการทำซ้ำ

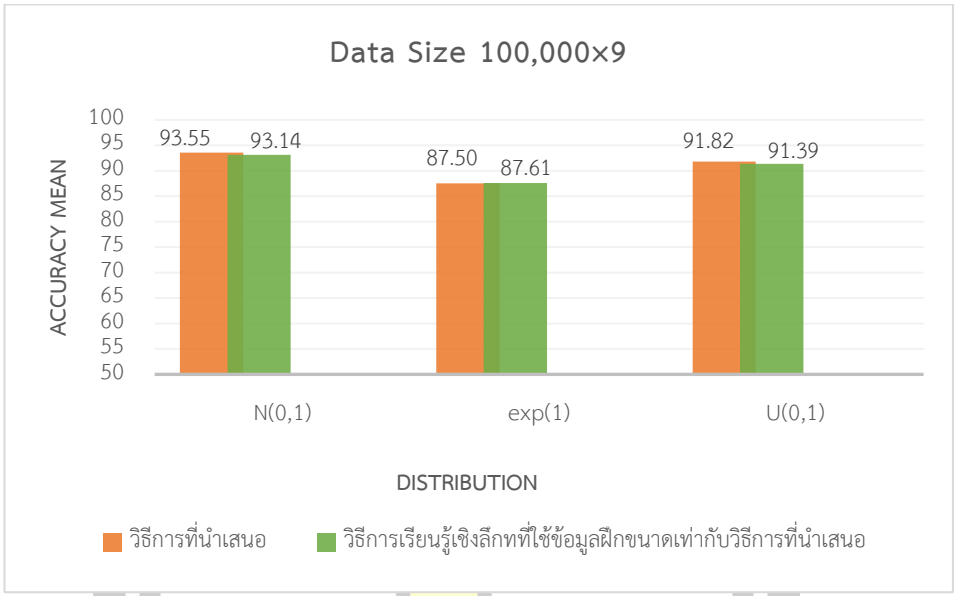


ภาพที่ 27 แสดงค่าความแม่นเฉลี่ยของชุดข้อมูลที่มีการแจกแจงเอกรูปใน 300 รอบของการทำซ้ำ

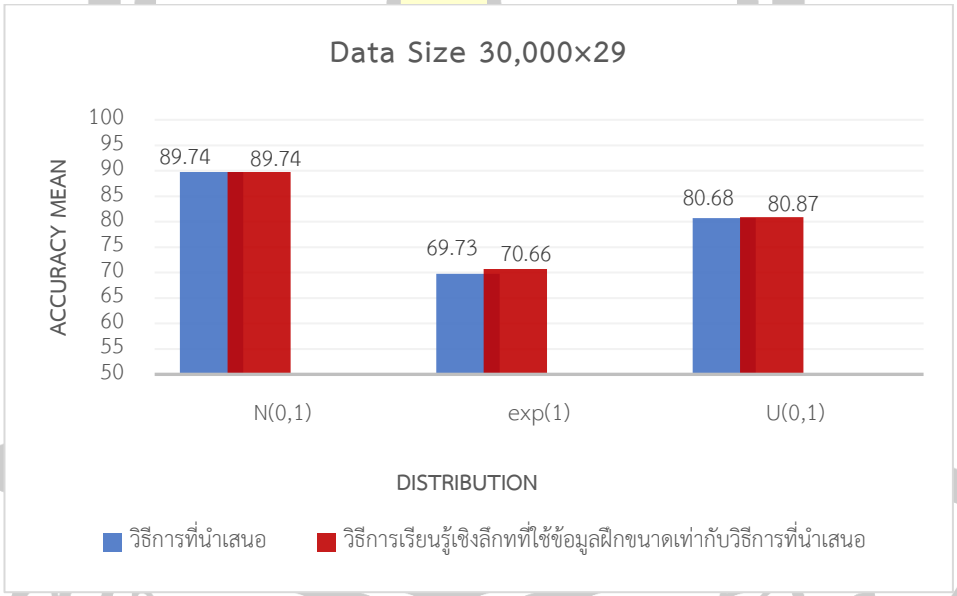
แสดงค่าความแม่นเฉลี่ยใน 300 รอบของการทำซ้ำโดยแบ่งตามขนาดข้อมูลของวิธีการที่นำเสนอและวิธีการเรียนรู้เชิงลึกที่ใช้ข้อมูลฝึกขนาดเท่ากับวิธีการที่นำเสนอของขนาดข้อมูลทั้ง 4 ขนาด ดังภาพที่ 28 - 31



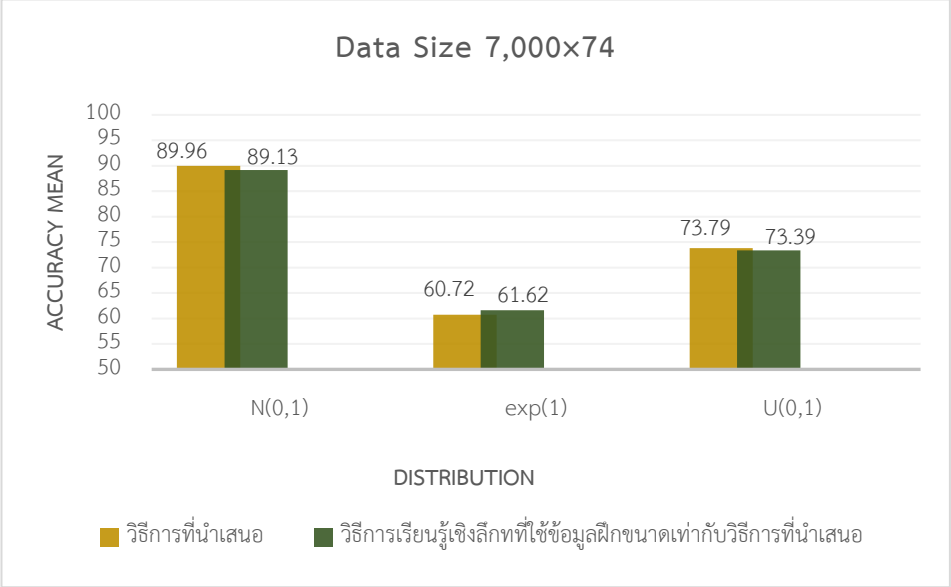
ภาพที่ 28 แสดงค่าความแม่นเฉลี่ยของทุกการแจกแจงที่ศึกษาใน 300 รอบของการทำซ้ำ



ภาพที่ 29 แสดงค่าความแม่นเฉลี่ยของทุกการแจกแจงที่ศึกษาใน 300 รอบของการทำซ้ำ

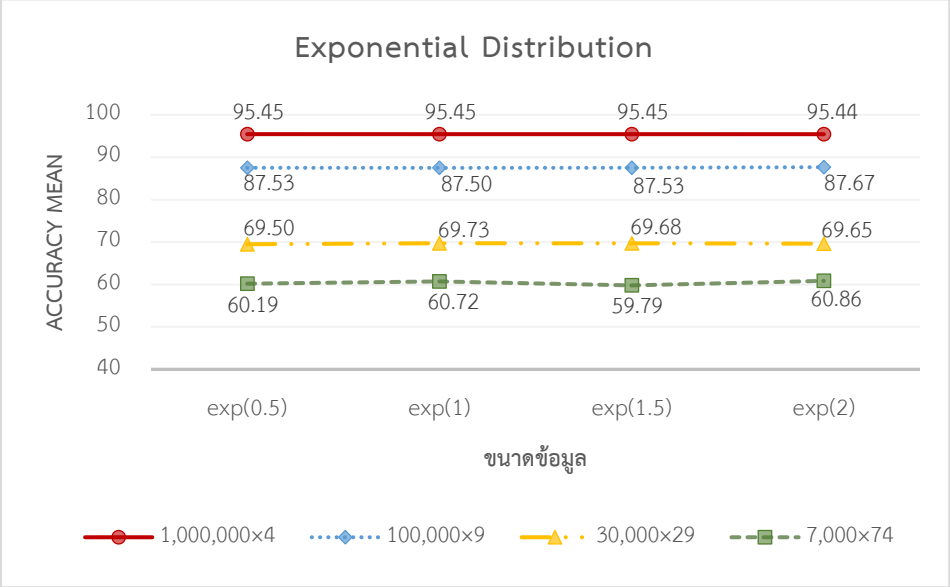


ภาพที่ 30 แสดงค่าความแม่นเฉลี่ยของทุกการแจกแจงที่ศึกษาใน 300 รอบของการทำซ้ำ



ภาพที่ 31 แสดงค่าความแม่นเฉลี่ยของทุกการแจกแจงที่ศึกษาใน 300 รอบของการทำซ้ำ

จากผลในตาราง 4 จะเห็นได้ว่าประสิทธิภาพความแม่นของวิธีการที่นำเสนอนั้นจะน้อยลงเมื่อข้อมูลมีการแจกแจงแบบเลขชี้กำลัง exp(1) ซึ่งเป็นการแจกแจงที่มีลักษณะเบ้ขวาจึงมีความเป็นไปได้ที่ความเบ้จะส่งผลต่อประสิทธิภาพความแม่น ดังนั้นผู้วิจัยจึงสนใจศึกษาว่าระดับความเบ้ที่แตกต่างจะส่งผลต่อประสิทธิภาพความแม่นหรือไม่ จึงทำการปรับพารามิเตอร์ของการแจกแจงแบบเลขชี้กำลังเป็น 4 แบบ คือ แบบที่ 1 exp(0.5) แบบที่ 2 exp(1) แบบที่ 3 exp(1.5) และแบบที่ 4 exp(2) ดังภาพที่ 32



ภาพที่ 32 แสดงค่าความแม่นเฉลี่ยของชุดข้อมูลที่มีการแจกแจงแบบเลขชี้กำลัง ใน 300 รอบของการทำซ้ำ

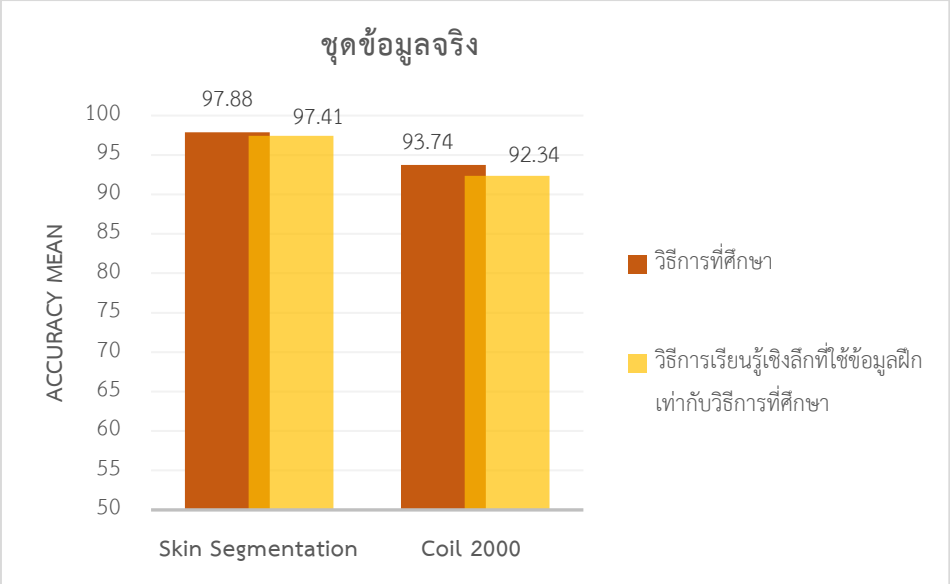
จากภาพที่ 32 พบว่าประสิทธิภาพความแม่นของชุดข้อมูลที่ Feature มีการแจกแจงแบบเลขชี้กำลังใน 300 รอบของการทำซ้ำ แต่ละขนาดที่ทำการศึกษาถึงแม้พารามิเตอร์จะเปลี่ยนไป ประสิทธิภาพความแม่นของชุดข้อมูลไม่ได้แตกต่างกัน

ตาราง 5 แสดงประสิทธิภาพของวิธีการที่นำเสนอจากชุดข้อมูลจริง โดยแสดงขนาดของข้อมูลฝึกและประสิทธิภาพความแม่นเมื่อเทียบกับวิธีการเรียนรู้เชิงลึกที่ใช้การสุ่มข้อมูลฝึกขนาดเท่ากับวิธีการที่นำเสนอ

การแจกแจงของข้อมูล	**ขนาดข้อมูล (N×Feature)**	**พารามิเตอร์**			**ขนาดข้อมูลฝึก**	**ประสิทธิภาพความแม่น**	
		สัดส่วน	**Km**	**RT**	**Case (%)**	**Mean of Accuracy %**	**Mean of AUC**
Skin Segmentation	245,057×3	0.10	25	5	183 (0.08%)	97.8810 [97.4139]	0.9863 [0.9794]
Coil2000	9,822×84	0.10	10	20	67 (0.68%)	93.7354 [92.3439]	0.9404 [0.9405]

หมายเหตุ : ค่าในวงเล็บ [ ] คือวิธีการเรียนรู้เชิงลึก

แสดงค่าความแม่นเฉลี่ยจากตารางที่ 5 ของวิธีการที่นำเสนอและวิธีการเรียนรู้เชิงลึกที่ใช้ข้อมูลฝึกขนาดเท่ากับวิธีการที่นำเสนอของข้อมูลจริงใน 1 รอบของการทำซ้ำ ดังภาพที่ 33



ภาพที่ 33 แสดงค่าความแม่นเฉลี่ยของชุดข้อมูลจริงใน 1 รอบของการทำซ้ำ

จากตาราง 5 แสดงประสิทธิภาพของวิธีการที่นำเสนอ โดยข้อมูลที่ใช้ในการศึกษาเป็นชุดข้อมูลจริงที่มีขนาดใหญ่มากจำนวน 2 ชุด ประกอบไปด้วยชุดข้อมูล Skin Segmentation โดยมีขนาดข้อมูล 245,057×3 จากฐานเก็บข้อมูล UCI และชุดข้อมูล Coil 2000 โดยมีขนาดข้อมูล 9,822×84 จากฐานเก็บข้อมูล KEEL ในแต่ละชุดข้อมูลจะแสดงการกำหนดพารามิเตอร์สัดส่วนในการสุ่มข้อมูลเพื่อใช้เป็นข้อมูลฝึก ในการศึกษานี้เพื่อลดเวลาในการประมวลผลจึงกำหนดสัดส่วนในการสุ่มที่ 0.10 หรือ 10% ในการนำไปใช้เพื่อเป็นข้อมูลฝึก สำหรับข้อมูล Skin Segmentation และข้อมูล Coil 2000 อีกทั้งยังแสดงการกำหนดกลุ่ม (Km=25 และ 10) ในการกำหนดกลุ่มสำหรับข้อมูล Skin Segmentation จำเป็นต้องเพิ่ม km ถึง 25 กลุ่ม เพื่อให้ได้ข้อมูลฝึกที่จะนำไปจำแนกประเภทมีข้อมูลจากทั้ง 2 Class โดยกำหนดการตัดข้อมูลที่มีค่ามากกว่าตำแหน่งเปอร์เซ็นต์ไทล์ที่ 95 รวมถึงการกำหนดรอบในการทำซ้ำ (RT=5 และ 20) ทั้งนี้รอบในการทำซ้ำเมื่อข้อมูลมีขนาดที่ไม่ใหญ่มากจนเกินไป เช่น ชุดข้อมูล Coil 2000 ที่มีขนาด 9,822×84 สามารถเพิ่มรอบในการทำซ้ำเพื่อให้ได้ประสิทธิภาพที่สูงขึ้น โดยใช้เวลาในการประมวลผลไม่นานนักและทำการกำหนดการตัดข้อมูลที่มีค่ามากกว่าตำแหน่งเปอร์เซ็นต์ไทล์ที่ 90

จากคอลัมน์ที่ 4 จะเห็นได้ว่าวิธีการที่นำเสนอสามารถลดขนาดของข้อมูลฝึกได้อย่างมาก โดยใช้ขนาดข้อมูลฝึกน้อยกว่า 1% ของจำนวนข้อมูลทั้งหมด แต่ยังคงให้ค่าความแม่นเฉลี่ยในการจำแนกประเภทสูงถึง 97.8810% และให้ค่า AUC สูงถึง 0.9863 สำหรับชุดข้อมูล Skin Segmentation ที่มีขนาด N จำนวนมาก และจำนวน Feature น้อย รวมถึงประสิทธิภาพความแม่นของชุดข้อมูล Coil 2000 ที่มี N จำนวนน้อย และจำนวน Feature มากขึ้น ซึ่งประสิทธิภาพความแม่นยังคงสูงกว่าวิธีการเรียนรู้เชิงลึกที่ใช้การสุ่มข้อมูลฝึกขนาดเท่ากับวิธีการที่นำเสนอ ทั้งนี้การเพิ่มการกำหนดจำนวนกลุ่มและการกำหนดรอบในการทำซ้ำ จะทำให้ได้ประสิทธิภาพในการจำแนกประเภทที่สูง แต่อาจส่งผลทำให้เวลาที่ใช้ในการประมวลผลช้าลง

ในส่วนนี้จะแสดงตัวอย่างการคำนวณหาค่าความแม่นเฉลี่ยและค่า AUC เฉลี่ย ใน 1 รอบของการทำซ้ำ โดยใช้ข้อมูล Skin Segmentation ซึ่งมีข้อมูลฝึกจำนวน 90 Case และมีข้อมูลทดสอบจำนวน 244,967 Case

ตาราง 6 แสดงค่าในตาราง Confusion Matrix ค่าความแม่นและค่า AUC จากการจำแนกประเภทด้วยวิธีที่ศึกษาจากจำนวนชั้นซ่อนและโหนดที่กำหนดทั้ง 27 กรณี

โหนด			TP	FP	TN	FN	Accuracy (%)	AUC
ชั้นซ่อน 1	ชั้นซ่อน 2	ชั้นซ่อน 3						
5	5	5	50727	7205	186936	6	97.0552	0.9814
5	5	10	50709	7711	186430	24	96.8412	0.9799
5	5	20	50697	4981	189160	36	97.9512	0.9868
5	10	5	50706	13164	180977	27	94.6131	0.9658
5	10	10	50630	1896	192245	103	99.1837	0.9941
5	10	20	50709	12097	182044	24	95.0501	0.9686
5	20	5	50660	5785	188356	73	97.6077	0.9844
5	20	10	50718	6271	187870	15	97.4330	0.9837
5	20	20	50714	10938	183203	19	95.5255	0.9716
10	5	5	50700	2576	191565	33	98.9346	0.9930
10	5	10	50700	4953	189188	33	97.9639	0.9869
10	5	20	50712	6139	188002	21	97.4844	0.9840
10	10	5	50721	7880	186261	12	96.7771	0.9796
10	10	10	50345	1684	192457	388	99.1539	0.9918
10	10	20	50722	6058	188083	11	97.5216	0.9843
10	20	5	50714	2771	191370	19	98.8606	0.9927
10	20	10	50690	1960	192181	43	99.1820	0.9945
10	20	20	50717	5016	189125	16	97.9451	0.9869
20	5	5	50710	3858	190283	23	98.4151	0.9898

โหนด			TP	FP	TN	FN	Accuracy (%)	AUC
ชั้นซ่อน 1	ชั้นซ่อน 2	ชั้นซ่อน 3						
20	5	10	50719	3791	190350	14	98.4461	0.9901
20	5	20	50733	2633	191508	0	98.9248	0.9932
20	10	5	50723	4787	189354	10	98.0410	0.9876
20	10	10	50718	2857	191284	15	98.8272	0.9925
20	10	20	50723	3941	190200	10	98.3865	0.9898
20	20	5	50647	3260	190881	86	98.6336	0.9908
20	20	10	50703	2438	191703	30	98.9921	0.9934
20	20	20	50714	2340	191801	19	99.0366	0.9938
ค่าความแม่นเฉลี่ย (%) และค่า AUC เฉลี่ย							97.8810	0.9863

จากตาราง 6 แสดงค่าในตาราง Confusion Matrix จากการจำแนกประเภทด้วยวิธีที่ศึกษาจากจำนวนชั้นซ่อนและโหนดที่กำหนดทั้ง 27 กรณี พบว่าเมื่อโหนดของแต่ละชั้นซ่อนเพิ่มมากขึ้น ทำให้ค่าความแม่นและค่า AUC สูง (> 0.97) หากหาค่าความแม่นเฉลี่ยและค่า AUC เฉลี่ยจาก 27 กรณี จะเห็นได้ว่าค่าความแม่นเฉลี่ยยังคงสูงถึง 97.8810% และค่า AUC เฉลี่ยคือ 0.9863 ใน 1 รอบของการทำซ้ำ

4.2 เปรียบเทียบประสิทธิภาพของการจำแนกประเภทและเวลากับวิธีการเรียนรู้เชิงลึกโดยใช้ข้อมูลฝึกขนาด 80% กับ 90%

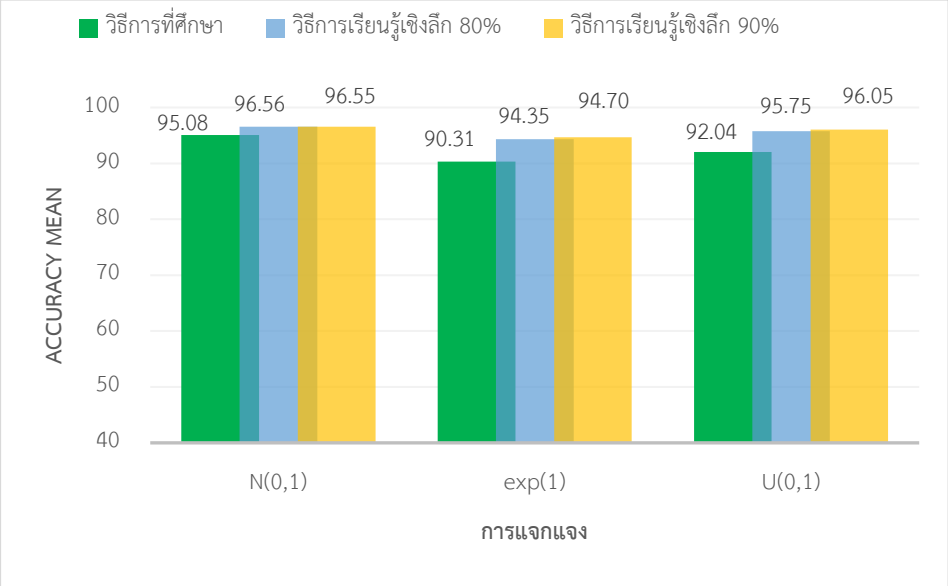
ในส่วนนี้จะแสดงให้เห็นถึงขนาดของข้อมูลฝึกที่ลดลงโดยเปรียบเทียบกับวิธีการเรียนรู้เชิงลึกโดยการสุ่มข้อมูลแบบ K-fold ซึ่งไม่ว่าข้อมูลจะมีมากน้อยเพียงใดค่า K ที่นิยมและถือว่าเป็นมาตรฐานคือ K=10 เพราะจะเหลือข้อมูลไว้สำหรับฝึกถึง 90% ในแต่ละรอบ ในการศึกษาครั้งนี้จะเปรียบเทียบประสิทธิภาพความแม่นและเวลาที่ใช้ในการประมวลผลทั้งหมดของการจำแนกประเภทระหว่างวิธีการเรียนรู้เชิงลึกกรณีที่ใช้ข้อมูลฝึก 80% กับ 90% และวิธีการที่นำเสนอ (**ตัวหนา**) ในตาราง 7 ใช้ชุดข้อมูลที่มีขนาด 7,000×74 เนื่องจากเป็นกรณีศึกษาพบว่ามีประสิทธิภาพในการจำแนกประเภทน้อยกว่าขนาดอื่นที่ใช้ในการศึกษา

ตาราง 7 แสดงประสิทธิภาพของวิธีการเรียนรู้เชิงลึกกรณีที่ใช้ข้อมูลฝึกขนาด 80% กับ 90% และวิธีการที่นำเสนอ

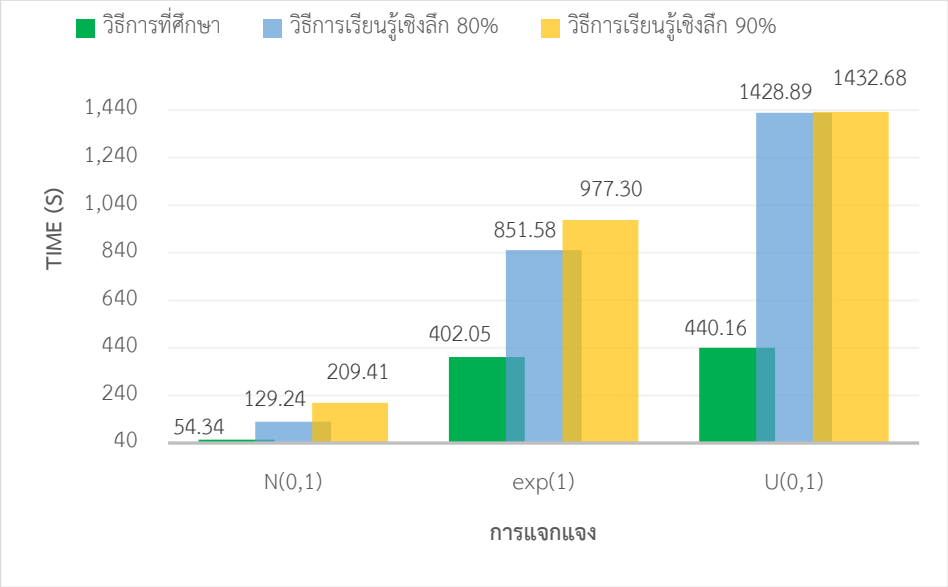
ข้อมูลและวิธีการ	**ขนาดข้อมูลฝึก**	**ประสิทธิภาพความแม่น**		**เวลา (วินาที)**
	Cases (%)	**Mean of Accuracy %**	**Mean of AUC**	
N(0,1); วิธีการเรียนรู้เชิงลึก	6,300 (90%)	96.5513	0.9413	209.41
	5,600 (80%)	96.5587	0.9425	129.25
วิธีการที่นำเสนอ	**524 (7.49%)**	**95.0804**	**0.9049**	**54.34**
สัดส่วน=0.30; Km=50; RT=30				
exp(1); วิธีการเรียนรู้เชิงลึก	6,300 (90%)	94.6989	0.9190	977.30
	5,600 (80%)	94.3476	0.9129	851.58
วิธีการที่นำเสนอ	**2,079 (29.70%)**	**90.3053**	**0.8612**	**402.05**
สัดส่วน=0.40; Km=200;RT=100				
U(0,1); วิธีการเรียนรู้เชิงลึก	6,300 (90%)	96.0540	0.9459	1432.68
	5,600 (80%)	95.7487	0.9429	1428.89
วิธีการที่นำเสนอ	**1,689 (24.13%)**	**92.0375**	**0.8819**	**440.16**
สัดส่วน=0.30; Km=200; RT=30				

หมายเหตุ: วิธีการที่นำเสนอ (**ตัวหนา**)

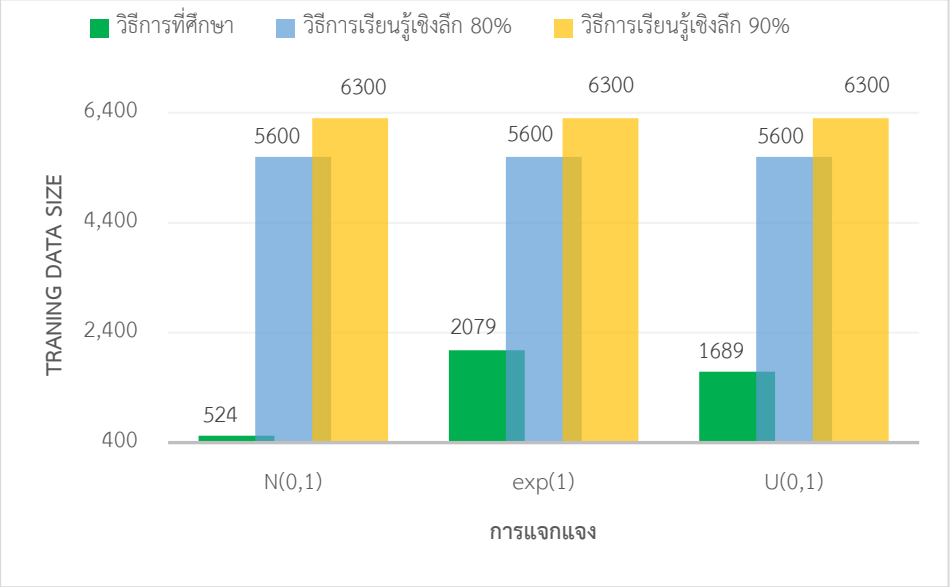
ในส่วนนี้จะเป็นการแสดงค่าความแม่น รวมถึงเวลาที่ใช้ในการประมวลผลและจำนวนข้อมูลฝึก ด้วยแผนภูมิแท่งระหว่างวิธีการเรียนรู้เชิงลึกที่ใช้ข้อมูลฝึกขนาด 80% กับ 90% และวิธีการที่นำเสนอของชุดข้อมูลที่มีการแจกแจงปรกติมาตรฐาน การแจกแจงแบบเลขชี้กำลัง รวมถึงชุดข้อมูลที่มีการแจกแจงเอกรูป และเป็นกรณีของชุดข้อมูลที่มีขนาด 7,000×74 ดังภาพที่ 34 - 36



ภาพที่ 34 แสดงค่าความแม่นเฉลี่ยของชุดข้อมูลขนาด 7,000×74



ภาพที่ 35 แสดงเวลาที่ใช้ในการประมวลผลของชุดข้อมูลขนาด 7,000×74



ภาพที่ 36 แสดงจำนวนข้อมูลฝึกของชุดข้อมูลขนาด 7,000×74

จากตาราง 7 แสดงประสิทธิภาพของวิธีการเรียนรู้เชิงลึกกรณีที่ใช้ข้อมูลฝึกขนาด 80% กับ 90% โดยข้อมูลที่ใช้ในการศึกษาเป็นข้อมูลที่มีขนาดใหญ่มากจากการสร้างข้อมูล (N × Feature) จำนวน 3 ชุด ประกอบไปด้วยชุดข้อมูลที่ Feature มีการแจกแจงปรกติมาตรฐาน ชุดข้อมูลที่ Feature มีการแจกแจงแบบเลขชี้กำลัง และชุดข้อมูลที่ Feature มีการแจกแจงเอกรูป พบว่าวิธีการเรียนรู้เชิงลึกที่ใช้ข้อมูลฝึกมากถึง 90% หรือ 6,300 Cases ของขนาดข้อมูลทั้งหมด มีประสิทธิภาพความแม่นเฉลี่ยในการจำแนกประเภทสูงในทุกการแจกแจง (> 94%) โดยค่าความแม่นเฉลี่ยและค่า AUC ของชุดข้อมูลที่ Feature มีการแจกแจงแบบเลขชี้กำลังให้ค่าน้อยกว่าชุดข้อมูลที่ Feature มีการแจกแจงปรกติมาตรฐานและชุดข้อมูลที่ Feature มีการแจกแจงเอกรูปเล็กน้อย ในส่วนของวิธีการที่นำเสนอเมื่อ Feature มีการแจกแจงปรกติมาตรฐาน ใช้ข้อมูลฝึกเพียง 7.49% ให้ค่าความแม่นเฉลี่ยสูงถึง 95.0804% แต่ใช้เวลาในการประมวลผลเพียง 54.34 วินาที ซึ่งน้อยกว่าวิธีการเรียนรู้เชิงลึกเกือบ 4 เท่า เมื่อ Feature มีการแจกแจงเอกรูปและการแจกแจงแบบเลขชี้กำลังจากตาราง 4 จะเห็นได้ว่าประสิทธิภาพในการจำแนกประเภท < 80% ในส่วนนี้จึงต้องใช้ข้อมูลฝึกเพิ่มมากขึ้นเพื่อให้ได้ประสิทธิภาพในการจำแนกประเภทที่สูง โดยทำการกำหนดสัดส่วนที่ 0.40 และ 0.30 กำหนด km ที่ 200 กลุ่ม กำหนด RT ที่ 100 และ 30 รอบ ตามลำดับการแจกแจง หากการแจกแจงเอกรูปกำหนดสัดส่วนและ RT ที่เท่ากับการแจกแจงปรกติมาตรฐานและการแจก

แจงแบบเลขชี้กำลัง จะได้ขนาดข้อมูลฝึกเป็น 1,504 Case ค่าความแม่นเฉลี่ยอยู่ที่ 87.9239% ค่า AUC เฉลี่ยคือ 0.8379 เวลาในการประมวลผล 215.87 วินาที ซึ่งผลลัพธ์โดยรวมยังสามารถเพิ่มประสิทธิภาพให้สูงขึ้นได้ โดยการปรับการกำหนดพารามิเตอร์ดังกล่าว ทั้งนี้เมื่อทำการปรับพารามิเตอร์แล้ว โดยจะใช้ข้อมูลฝึกประมาณ 20% ถึง 30% ซึ่งให้ค่าความแม่นเฉลี่ยที่สูงขึ้น (> 90%) โดยเวลาที่ใช้ในการประมวลผลนั้นน้อยกว่าวิธีการเรียนรู้เชิงลึกอย่างเห็นได้ชัด

บทที่ 5
สรุปผล อภิปรายผล และข้อเสนอแนะ

จากผลการวิเคราะห์ด้วยวิธีการเรียนรู้เชิงลึกโดยใช้ข้อมูลฝึกที่ผ่านคุณสมบัติของวิธีเคมีนและการตรวจหาค่าผิดปกติ รวมถึงผลการวิเคราะห์ด้วยวิธีการเรียนรู้เชิงลึกแบบเดิมที่ใช้ข้อมูลฝึกขนาดต่าง ๆ เช่น ข้อมูลฝึกขนาด 80% กับ 90% เป็นต้น สามารถสรุปผลการศึกษา อภิปรายผล และข้อเสนอแนะได้ดังนี้

5.1 สรุปผล

5.2 อภิปรายผล

5.3 ข้อเสนอแนะ

5.1 สรุปผลการวิจัย

จากการศึกษาการจำแนกข้อมูลขนาดใหญ่มากโดยใช้การวิเคราะห์กลุ่มด้วยวิธีเคมีนและวิธีการเรียนรู้เชิงลึก เพื่อลดปัญหาของเวลาในการประมวลผลซึ่งต้องใช้เวลานานและต้องใช้ข้อมูลฝึกเป็นจำนวนมาก เพื่อยังคงประสิทธิภาพความแม่นที่สูง จึงทำการลดขนาดข้อมูลฝึกด้วยการรวมเทคนิคการจัดกลุ่มของวิธีเคมีนและวิธีการเรียนรู้เชิงลึก จากผลการศึกษาพบว่าวิธีการที่นำเสนอสามารถลดขนาดข้อมูลฝึกได้อย่างมาก โดยเฉพาะอย่างยิ่งในกรณีของชุดข้อมูลที่มีขนาด N จำนวนมาก และจำนวน Feature น้อย สามารถใช้ข้อมูลฝึกน้อยกว่า 1% ของจำนวนข้อมูลทั้งหมด แต่ยังคงให้ค่าความแม่นเฉลี่ยและค่า AUC เฉลี่ยสูงมาก เมื่อเปรียบเทียบกับวิธีการเรียนรู้เชิงลึกที่ใช้การสุ่มข้อมูลฝึกขนาดเท่ากับวิธีการที่นำเสนอ จากการสร้างข้อมูล (N × Feature) ในชุดข้อมูลที่ Feature มีการแจกแจงปรกติมาตรฐานและการแจกแจงเอกรูป วิธีการที่นำเสนอนั้นมีประสิทธิภาพความแม่นเฉลี่ยที่สูงกว่าอีกทั้งยังให้ค่า AUC อยู่ในเกณฑ์ที่ตัวแบบทำงานได้ดีถึงดีมาก ในขณะที่ชุดข้อมูลที่ Feature มีการแจกแจงแบบเลขชี้กำลังมีประสิทธิภาพความแม่นเฉลี่ยที่เทียบเท่ากันแต่ยังคงให้ค่า AUC ที่อยู่ในเกณฑ์มาตรฐานสำหรับตัวแบบส่วนใหญ่ เมื่อพิจารณาผลที่ได้จากวิธีการที่นำเสนอเทียบกับวิธีการเรียนรู้เชิงลึกที่ใช้ข้อมูลฝึก 80% และ 90% ของข้อมูลทั้งหมด พบว่ามีประสิทธิภาพความแม่นในการจำแนกประเภทสูงในทุกการแจกแจง (> 94%) โดยเฉพาะอย่างยิ่งกรณีที่ Feature ข้อมูลมีการแจกแจงปกติเมื่อใช้ข้อมูลฝึก 90% หรือ 6,300 Cases ของขนาดข้อมูลทั้งหมด ให้ค่า

ความแม่นเฉลี่ยในการจำแนกประเภทสูงมาก (> 96%) โดยใช้เวลาในการประมวลผล 209.41 วินาที ในขณะที่วิธีการที่นำเสนอใช้ข้อมูลฝึกน้อยกว่า 30% ของข้อมูลทั้งหมด แต่ยังคงประสิทธิภาพความแม่นในการจำแนกประเภทสูง (> 90%) ซึ่งเห็นได้ชัดในกรณีชุดข้อมูลที่ Feature มีการแจกแจงปรกติมาตรฐานโดยใช้ข้อมูลฝึกเพียง 7.49% หรือ 524 Case ของขนาดข้อมูลทั้งหมด แต่ยังคงให้ค่าความแม่นเฉลี่ยในการจำแนกประเภทสูง (> 95%) ซึ่งเกือบเท่าวิธีการเรียนรู้เชิงลึกที่ใช้ข้อมูลฝึก 90% อีกทั้งยังใช้เวลาในการประมวลผลเพียง 54.34 วินาที ซึ่งน้อยลงเกือบ 4 เท่า เมื่อเทียบกับวิธีการเรียนรู้เชิงลึกที่ใช้ข้อมูลฝึก 90% ในกรณีของชุดข้อมูลที่ Feature มีการแจกแจงแบบเลขชี้กำลัง ก็ยังคงประสิทธิภาพความแม่นในการจำแนกประเภทสูง (> 92%) และใช้เวลาในการประมวลผลน้อยกว่าวิธีการเรียนรู้เชิงลึกที่ใช้ข้อมูลฝึก 90% และในกรณีชุดข้อมูลที่ Feature มีการแจกแจงเอกรูป: exp(1) โดยให้ค่าความแม่นเฉลี่ยในการจำแนกประเภทลดลงเล็กน้อย (> 90%) ซึ่งไม่ต่างกับ exp(0.5) exp(1.5) และ exp(2) ที่มีพารามิเตอร์ต่างกัน

5.2 อภิปรายผลการวิจัย

จากผลการวิจัยของวิธีการเรียนรู้เชิงลึกโดยใช้ข้อมูลฝึกที่ผ่านคุณสมบัติของวิธีเคมีนและการตรวจหาค่าผิดปกติ ในส่วนของการเปรียบเทียบประสิทธิภาพวิธีการที่นำเสนอกับวิธีการเรียนรู้เชิงลึกที่ใช้การสุ่มข้อมูลฝึกขนาดเท่ากับวิธีการที่นำเสนอทั้งจากข้อมูลที่สร้างขึ้นและข้อมูลจริง พบว่าวิธีการที่นำเสนอสามารถลดขนาดของข้อมูลฝึกได้อย่างมากในทุกการแจกแจงที่ศึกษา โดยใช้ขนาดข้อมูลฝึกน้อยกว่า 1% ของจำนวนข้อมูลทั้งหมด ในกรณีที่ชุดข้อมูลมีขนาด N จำนวนมาก และจำนวน Feature น้อย (1,000,000×4) ให้ค่าความแม่นเฉลี่ยสูงมาก (> 95%) และให้ค่า AUC สูงมาก (> 0.90) ในทุกการแจกแจง เมื่อกรณีที่จำนวน Feature มากขึ้น และขนาด N จำนวนน้อยลง (7,000×74) ค่าความแม่นเฉลี่ยและค่า AUC ของการจำแนกประเภทจะน้อยลงแต่ยังคงมีค่าสูงกว่าผลจากวิธีการเรียนรู้เชิงลึกที่ใช้การสุ่มข้อมูลฝึกขนาดเท่ากับวิธีการที่นำเสนอ ซึ่งสอดคล้องกับงานวิจัยของ [3] ที่ทำการลดขนาดข้อมูลฝึก โดยจะมีประสิทธิภาพในการจำแนกประเภทสูงในกรณีข้อมูลมีขนาด N จำนวนมาก และจำนวน Feature น้อย เมื่อข้อมูลมีขนาด N จำนวนน้อยลง และจำนวน Feature มากขึ้น จะทำให้ประสิทธิภาพในการจำแนกประเภทลดลงเช่นเดียวกับวิธีการที่นำเสนอ

หากพิจารณาตามการแจกแจงพบว่า ประสิทธิภาพความแม่นของวิธีการที่นำเสนอนั้นจะสูง เมื่อ Feature มีการแจกแจงปรกติมาตรฐาน (> 90%) และการแจกแจงเอกรูป (> 75%) และจะน้อยลงเมื่อ Feature มีการแจกแจงแบบเลขชี้กำลัง (< 70%) แต่ทั้งนี้ยังให้ค่าความแม่นเทียบเท่ากับวิธีการเรียนรู้เชิงลึกที่ใช้การสุ่มข้อมูลฝึกเท่ากับวิธีการที่นำเสนอ ซึ่งในการศึกษาครั้งนี้ทำการสร้างข้อมูล (N × Feature) และได้ทำการกำหนด Target เป็น 3 Class เนื่องจากไม่สามารถแบ่ง Class ได้ชัดเจน ทำให้เมื่อมี Feature มากยิ่งขึ้น ประสิทธิภาพความแม่นในการจำแนกประเภทจึงน้อยลงทั้งวิธีการที่นำเสนอและวิธีการเรียนรู้เชิงลึก [39]

ในการกำหนดจำนวนของชั้นซ่อนที่ใช้ในวิธีการเรียนรู้เชิงลึก ผู้วิจัยได้ทดลองกำหนดจำนวนชั้นซ่อนเท่ากับ 2 3 5 10 15 20 25 30 35 40 45 และ 50 และพบว่าประสิทธิภาพในการจำแนกประเภทสูงในทุกจำนวนชั้นซ่อนที่กำหนด แต่ยิ่งกำหนดจำนวนชั้นซ่อนมากก็จะยิ่งใช้เวลาในการประมวลผลนานมาก นอกจากนี้ยังพบว่าหากกำหนดชั้นซ่อนมากกว่า 3 ชั้น ค่าความแม่นเฉลี่ยในการจำแนกประเภทไม่ได้แตกต่างจากการกำหนดจำนวนชั้นซ่อนเท่ากับ 3 มากนัก จึงได้กำหนดจำนวนชั้นซ่อนที่ใช้ในการศึกษาครั้งนี้เท่ากับ 3 ชั้นซ่อน นอกจากนี้ผู้วิจัยได้ลองกำหนดจำนวนโหนดของแต่ละชั้นซ่อนเป็น 2 3 5 10 15 20 25 30 35 40 45 และ 50 เช่นกัน พบว่าประสิทธิภาพความแม่นในการจำแนกประเภทสูงและไม่แตกต่างกันมากนัก แต่การกำหนดโหนดในแต่ละชั้นซ่อนสูง ๆ จะทำให้เวลาในการประมวลผลนานขึ้น ผู้วิจัยจึงสนใจศึกษาค่าของชั้นซ่อนและโหนดที่อยู่ในช่วง 5 ถึง 20 โดยทำการกำหนดชั้นซ่อนเป็น 3 ชั้นซ่อน โดยในแต่ละชั้นซ่อนกำหนดโหนดเป็น 5 10 และ 20 ซึ่งรวมเป็น 27 กรณี และโหนดทั้ง 27 กรณี มีความเหมาะสมกับข้อมูลขนาดใหญ่มากโดยให้ค่าความแม่นเฉลี่ยในการจำแนกประเภทสูงดังตารางผลการศึกษา ในส่วนของการกำหนดสัดส่วนในการสุ่มข้อมูลฝึก กำหนดการแบ่งกลุ่มของข้อมูลฝึก (Km) และการกำหนดรอบในการทำซ้ำ (RT) จากผลการศึกษาจะเห็นได้ว่าเมื่อ Feature มีการแจกแจงแบบเลขชี้กำลัง การกำหนดสัดส่วนในการสุ่มรวมถึง Km และ RT จะกำหนดมากกว่าการแจกแจงอื่น เพื่อให้ค่าความแม่นเฉลี่ยในการจำแนกประเภทสูง ซึ่งส่งผลทำให้เวลาที่ใช้ในการประมวลผลนานมากขึ้น แต่ทั้งนี้ยังคงใช้เวลาประมวลผลน้อยกว่าวิธีการเรียนรู้เชิงลึกที่ใช้ข้อมูลฝึก 80% กับ 90% ของข้อมูลทั้งหมด

จากผลการศึกษาจะเห็นได้ว่าวิธีการที่นำเสนอเหมาะสำหรับกรณีที่ชุดข้อมูลมีขนาด N จำนวนมาก และจำนวน Feature น้อย โดยเป็นชุดข้อมูลที่ Feature มีการแจกแจงปรกติมาตรฐาน รวมทั้งการแจกแจงเอกรูป หากชุดข้อมูลมีขนาด N จำนวนน้อย และจำนวน Feature มาก ยังถือว่าอยู่ในเกณฑ์มาตรฐานสำหรับตัวแบบส่วนใหญ่ เว้นแต่ชุดข้อมูลที่ Feature มีการแจกแจงแบบเลขชี้

กำลังซึ่งต้องเพิ่มจำนวนข้อมูลฝึกให้มากขึ้น เพื่อให้ยังคงมีประสิทธิภาพในการจำแนกประเภทที่สูง ในเรื่องของประสิทธิภาพในการจำแนกประเภทถึงแม้วิธีการที่นำเสนอจะมีประสิทธิภาพที่น้อยกว่าวิธีการเรียนรู้เชิงลึก แต่ถ้าหากพิจารณาในเรื่องของเวลาในการประมวลผล โดยเฉพาะชุดข้อมูลที่มีขนาด N จำนวนมาก และจำนวน Feature น้อย โดยเฉพาะเมื่อ Feature มีการแจกแจงปรกติ วิธีการที่นำเสนอก็เป็นทางเลือกที่น่าสนใจเพราะนอกจากลดระยะเวลาในการประมวลผลลงแล้ว ประสิทธิภาพในการจำแนกประเภทก็ยังสูงมากอีกด้วย

5.3 ข้อเสนอแนะ

ข้อเสนอแนะที่คาดว่าจะสามารถพัฒนางานวิจัยนี้ต่อไปได้อีก และเป็นประโยชน์สำหรับนักวิจัยท่านอื่นที่สนใจในด้านนี้ต่อไป

5.3.1 ข้อเสนอแนะจากการวิจัย

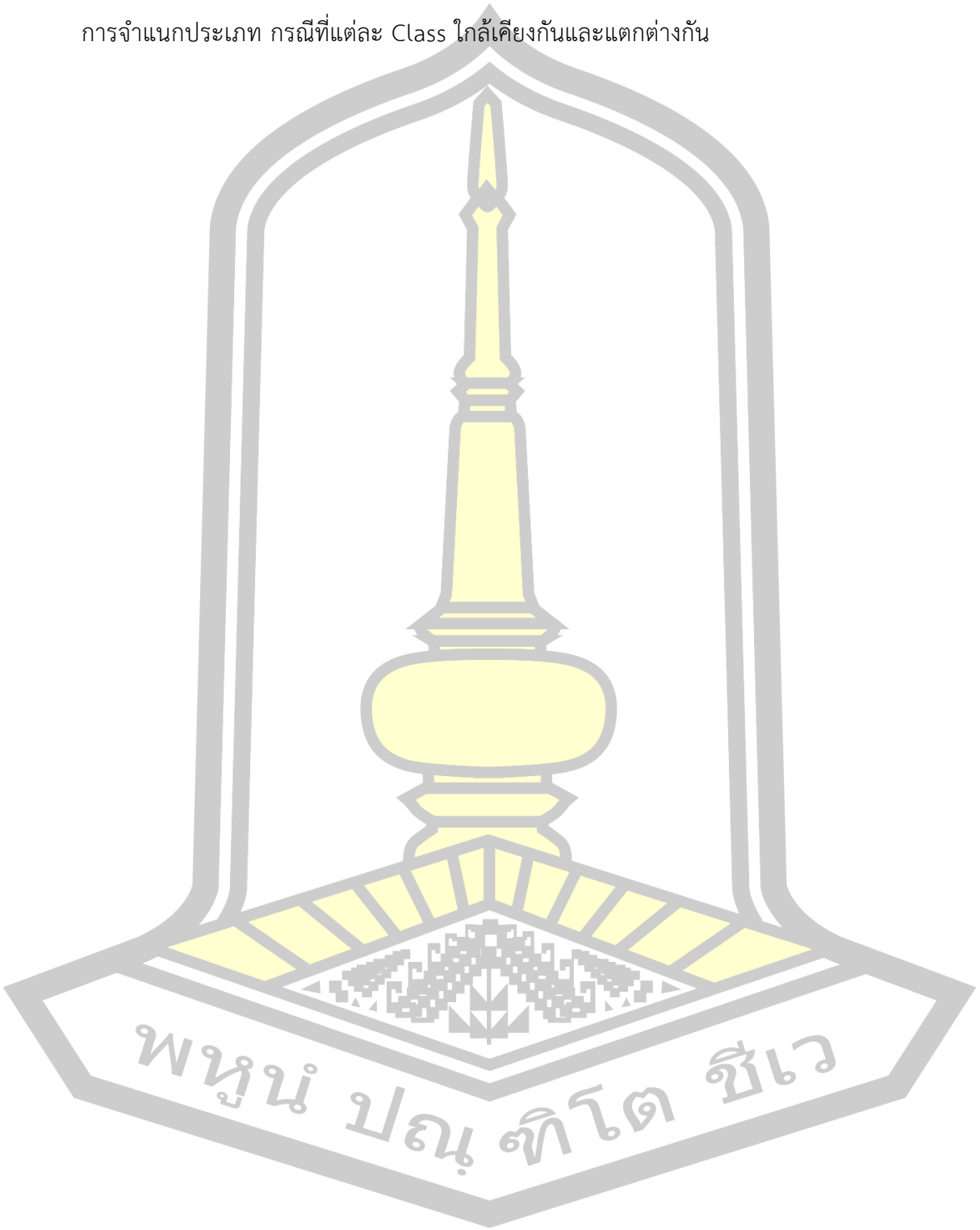
จากผลการวิจัยเรื่องการจำแนกข้อมูลขนาดใหญ่มากโดยใช้การวิเคราะห์กลุ่มด้วยวิธีเคมีนและวิธีการเรียนรู้เชิงลึก ในการกำหนดพารามิเตอร์การกำหนดจำนวนกลุ่ม (Km) รวมถึงรอบในการทำซ้ำของคุณสมบัติวิธีเคมีน (RT) และการทำซ้ำทั้งหมดนั่นคือทำซ้ำทั้งวิธีเคมีนและวิธีการเรียนรู้เชิงลึก ในการกำหนดพารามิเตอร์เหล่านี้ส่งผลต่อค่าความแม่นในการจำแนกประเภทและค่า AUC อีกทั้งยังส่งผลต่อเวลาในการประมวลผลอีกด้วย ทั้งนี้การกำหนดพารามิเตอร์ควรพิจารณาตามความเหมาะสมทั้งเรื่องเวลาที่ใช้ในการประมวลผล และประสิทธิภาพในการจำแนกประเภทร่วมด้วย

5.3.2 ข้อเสนอแนะเพื่อการวิจัยครั้งต่อไป

1) ในส่วนของการลดขนาดข้อมูลฝึกโดยใช้การวิเคราะห์กลุ่มด้วยวิธีเคมีนสามารถใช้เป็นวิธี k-means++ [40] เพื่อลดเวลาในการประมวลผลได้เร็วขึ้น แต่เพื่อยังคงประสิทธิภาพความแม่นที่สูง ทั้งนี้ข้อมูลก็ยังคงต้องมีขนาดใหญ่มากเช่นกัน

2) วิธีการเรียนรู้เชิงลึกที่ทำการคำนวณผ่าน ANN ในการศึกษาครั้งนี้ได้ทำการกำหนด Target เป็น 3 Class โดยได้จากค่าเฉลี่ยของ Feature ในแต่ละ Case ซึ่งเป็นค่าที่ใกล้เคียงกันหรือแทบไม่ต่างกัน จึงค่อนข้างยากต่อการจำแนกประเภท หากนักวิจัยท่านอื่นสนใจวิจัยเรื่องนี้สามารถลองปรับ Target ที่น้อยกว่า 3 Class เพื่อประสิทธิภาพความแม่นที่อาจจะสูงขึ้น

4) ในการแบ่ง Class สามารถเพิ่มการแบ่ง Class เพื่อศึกษาประสิทธิภาพของการจำแนกประเภท กรณีที่แต่ละ Class ใกล้เคียงกันและแตกต่างกัน



บรรณานุกรม

บรรณานุกรม

[1] สายชล สินสมบูรณ์ทอง. “การทำเหมืองข้อมูล (Data Mining)”. 2561. [ออนไลน์]. [สืบค้นเมื่อ 14 ธันวาคม 2562]; ได้จาก : http://www.thaiall.com/dm/index.html

[2] สุพรรณ ฟ้ายง. “เตรียมข้อมูลเพื่อสร้าง Model”. 2560. [ออนไลน์]. [สืบค้นเมื่อ 25 สิงหาคม 2562]; ได้จาก : http://codeonthehill.com/machine-learning-4-data-set/

[3] T. Tang, S. Chen, M. Zhao, W. Huang, and J. Luo, “Very large-scale data classification based on K-means clustering and multi-kernel SVM”. 2018, Soft Computing, 23(11), 3793-3801. doi:10.1007/s00500-018-3041-0.

[4] กฤศณัฏฐ์ บุญเกียรติพงษ์. “การประยุกต์นิวรอลเน็ตเวิร์กหลายโครงข่ายบนข้อมูลขนาดใหญ่”. [วิทยาศาสตรมหาบัณฑิต]. จุฬาลงกรณ์มหาวิทยาลัย; 2554.

[5] กานกวิญจน์ โค้วสีหวัฒน์. “แบบจำลองทำนายผลคำตัดสินและประเด็นในคดีอาญาที่เรียนรู้จากคำพิพากษาศาลฎีกาไทย โดยใช้เทคนิคการเรียนรู้เชิงลึก”. [วิทยาศาสตรมหาบัณฑิต]. จุฬาลงกรณ์มหาวิทยาลัย; 2561.

[6] อานนท์ ศักดิ์วรวิชญ์. “เทคโนโลยีการจัดการข้อมูลขนาดใหญ่”. 2560. [ออนไลน์]. [สืบค้นเมื่อ 26 สิงหาคม 2562]; ได้จาก : https://mgronline.com/daily/detail/959000 0063416

[7] Nessessence. “Deep Learning”. 2018. [ออนไลน์]. [สืบค้นเมื่อ 2 ตุลาคม 2562]; ได้จาก : https://www.thaiprogrammer.org/2018/12/deeplearning

[8] Namjatturas. “ทำความรู้จัก AI, Machine Learning, Deep Learning”. 2019. [ออนไลน์]. [สืบค้นเมื่อ 1 ตุลาคม 2562]; ได้จาก : https://techsauce.co/tech-and-biz/aimachine-learning-deep-learning-differences

[9] ตุลยา วุฑฒิปรีชา. “Data Analytic โอกาสสำหรับธุรกิจในยุค Digital”. 2562. [ออนไลน์]. [สืบค้นเมื่อ 22 สิงหาคม 2562]; ได้จาก : https://webcache.googleusercontent.com/search?q=cache:t1ERl1X_IYJ:https://www.gsb.or.th/getattachment/

[10] วราวุธ วุฒิวณิชย์. “Artificial Neural Networks (ANNs)”. 2545. [ออนไลน์]. [สืบค้นเมื่อ 20 สิงหาคม 2562]; ได้จาก : http://irre.ku .ac.th/slideshow/pdf/22.pdf

[11] Minaphinant. “Deep Learning คืออะไร”. 2018. [ออนไลน์]. [สืบค้นเมื่อ 29 สิงหาคม 2562]; ได้จาก : https://blog.finnomena.com/deep-learning

[12] KDD. “Data Mining”. 2007. [ออนไลน์]. [สืบค้นเมื่อ 14 ธันวาคม 2562]; ได้จาก : http://thailand-kdd.blogspot.com/2007/06/data-mining-thai.html

[13] กษิดิศ สตางค์มงคล. “Machine Learning 101 สร้างตัวแบบด้วย Excel”. 2560. [ออนไลน์]. [สืบค้นเมื่อ 25 กันยายน 2563]; ได้จาก : https://medium.com/@kasidissatangmong kol/machine-learning-101

[14] พีรัชต์ ลิ้มกรโชติวัฒน์. “เริ่มเรียน Machine Learning 0–100 zero to Mr.incredible (Introduction)”. 2561. [ออนไลน์]. [สืบค้นเมื่อ 15 กันยายน 2563]; ได้จาก : https://medium.com/mmp-li/machine-learning-0-100-introduction-1c58e516bfcd

[15] กัลยา วานิชย์บัญชา. “ประเภทของการจัดกลุ่มหรือแบ่งกลุ่ม (Cluster Analysis)”. 2560. [ออนไลน์]. [สืบค้นเมื่อ 15 สิงหาคม 2562]; ได้จาก : http://www.pratya.nuankaew.com/wp-content/uploads/2017/10/cluster-analysis.pdf

[16] จิตรลดา ทองอันตัง, สุขสมพร อโนไท. “การวิเคราะห์กลุ่ม (Cluster Analysis)”.2562. [ออนไลน์]. [สืบค้นเมื่อ 17 สิงหาคม 2562]; ได้จาก : https://rci2010.files.wordpress.com

[17] ณัฐวุฒิ ทองจอ. “Machine Learning รู้จักการจำแนกประเภทข้อมูลด้วย k-Nearest Neighbors”. 2560. [ออนไลน์]. [สืบค้นเมื่อ 15 กันยายน 2563]; ได้จาก : https://www.babelcoder.com/blog/articles/k-nearest-neighbors

[18] KimMS. “Robust, Scalable Anomaly Detection for Large Collections of Images”. International Conference on Social Computing (SocialCom), 2013, pp 1054–1058.

[19] Aheadasia. “AI คืออะไร ต่างจาก Machine Learning และ Deep Learning อย่างไร”. 2018. [ออนไลน์]. [สืบค้นเมื่อ 16 ธันวาคม 2562]; ได้จาก : https://www.Youtube.com/watch?v=45c--4XxGZs

[20] ศรัณย์ คชเสถียร. “Machine Learning: ANN”. 2560. [ออนไลน์]. [สืบค้นเมื่อ 25 สิงหาคม 2562]; ได้จาก : https://medium.com/@sarankhotsathian/machine-learning-ann

[21] Sivadee. “MACHINE LEARNING – นิยามและตัวอย่างการใช้งาน”. 2018. [ออนไลน์]. [สืบค้นเมื่อ 5 ตุลาคม 2562]; ได้จาก : https://www.toolmakers.co/machine-learning

[22] วรรณพงษ์ ภัททิยไพบูลย์. “Overfitting และ Underfitting”. 2561. [ออนไลน์]. [สืบค้นเมื่อ 30 สิงหาคม 2562]; ได้จาก https://the-ai-midnight.blogspot.com/2018/12/overfitting-underfitting.html

[23] ณัฐธัญ วิโรจน์บุญเกียรติ. “การระบุตัวคนขับรถโดยใช้ฮิสโทแกรมและโครงข่ายประสาทเทียมจากข้อมูลความเร่ง”. [วิศวกรรมศาสตรมหาบัณฑิต]. จุฬาลงกรณ์มหาวิทยาลัย; 2560.

[24] นัธทวัฒน์ รักสะอาด. “กระบวนการเพื่อการค้นคืนรูปภาพลายผ้าไหมที่มีกลุ่มตัวอย่างน้อย”. [ปริญญาวิทยาศาสตรมหาบัณฑิต]. มหาวิทยาลัยมหาสารคาม; 2561.

[25] วิรัช หิรัญ, ฐิตาภรณ พอบุตรด. “แบบจำลองการดึงดูดการเดินทางโดยใชขอมูลเครือขายสังคมออนไลนและการเรียนรูเชิงลึก”. วารสารศรีปทุมปริทัศน์ ฉบับวิทยาศาสตร์และเทคโนโลยี, 2561.

[26] ณัฏฐณิชา ยงยิ่ง. “การประยุกต์ใช้เทคโนโลยีการเรียนรู้เชิงลึกในการจำแนกข้อมูลถนนจากภาพถ่าย Drone เพื่อการสำรวจถนนในเขตชนบท”. [ปริญญาวิทยาศาสตรบัณฑิตสาขาวิชาภูมิศาสตร์]. มหาวิทยาลัยนเรศวร; 2562.

[27] นันทิพัฒน์ พลบดี. “การแยกพื้นที่บาดแผลจากภาพถ่ายด้วยการเรียนรู้เชิงลึกและการขยายข้อมูลแบบต่าง ๆ”. [วิทยาศาสตรมหาบัณฑิต]. มหาวิทยาลัยศิลปากร; 2561.

[28] นิภาพร ชุติมันต์, บังอร กุมพล, ศิริลักษณ์ เจิมจิตต์พรชัย, กฤตยา แสนคักดิ์, จันทร์เจริญ รักษ์มณี, นงนุช แสงสุระ, ปัทมวดี นันทนาเนตร์, โรจนี หอมชาลี, อรุณ แก้วมั่น. “ระเบียบวิธีการทางสถิติ”. มหาสารคาม : หจก.อภิชาติการพิมพ์; 2554.

[29] Coraline. (2018). “Machine Learning ไม่ได้มีแค่ Deep Learning”. 2018. [ออนไลน์]. [สืบค้นเมื่อ 16 ธันวาคม 2562]; ได้จาก : https://www.coraline.co.th/single-post/2018/09/09/Machine-Learning-is-not-just-Deep-Learning

[30] M. Sokolova, and G. Lapalme, “A systematic analysis of performance measures for classification tasks”. 2019, 427-437. doi:10.1016/j.ipm.2009.03.002.

[31] A. Muhammad, S. Dae-Hee, K. Sang-Hee, and N. Soon-Ryul, “An Accurate CT Saturation Classification Using a Deep Learning Approach Based on Unsupervised Feature Extraction and Supervised Fine-Tuning Strategy”. 2017, 10, 1830. doi:10.3390/en10111830

[32] AI. Marqués, V. García, and JS. Sánchez, “On the Suitability of Resampling Techniques for the Class Imbalance Problem in Credit Scoring”. 2013, 64, 1060-1070. doi:10.1057/jors.2012.120

[33] D.W. Hosmer, and S. Lemeshow, “Applied logistic regression”. A Wiley-Interscience Publication: p.162; 2013.

[34] กองพัท พิชากุล. “เรื่องการวัดประสิทธิภาพ”. 2558. [ออนไลน์]. [สืบค้นเมื่อ 27 สิงหาคม 2562]; ได้จาก : https://medium.com/o-v-e-r-f-i-t-t-e-d

[35] วนิดา พงษ์สงวน, ทิพยา ถินสูงเนิน, มาโนช ถินสูงเนิน. “การพัฒนาแบบจำลองปัจจัยที่มีผลต่อการเป็นโรคเบาหวานด้วยเทคนิคต้นไม้ตัดสินใจ”. 2562.

[36] เอกสิทธิ์ พัชรวงศ์ศักดา. “การแบ่งข้อมูลเพื่อทดสอบประสิทธิภาพของตัวแบบ”. 2558. [ออนไลน์]. [สืบค้นเมื่อ 17 กันยายน 2563]; ได้จาก : https://th.linkedin.com/pulse/-eakasit-pacharawongsa

[37] ณัฏฐา กาญจนขุนดี. “Big Data ข้อมูลขนาดใหญ่คืออะไร ทำงานอย่างไร และนำไปใช้อะไรได้บ้าง”. 2561. [ออนไลน์]. [สืบค้นเมื่อ 20 สิงหาคม 2562]; ได้จาก : https ://www.khundee.com.

[38] กษิดิศ สตางค์มงคล. “อธิบาย 10 Metrics พื้นฐานสำหรับวัดผลตัวแบบ Machine Learning”. 2561. [ออนไลน์]. [สืบค้นเมื่อ 25 สิงหาคม 2562]; ได้จาก : https://datarockie.com/2019/03/30/top-ten-machine-learning-metrics

[39] พีรธรรม วิริยธรรมภูมิ. “วิธีการเลือกฐานหลักสำหรับเทคนิคการลดมิติ”. [ปริญญาวิศวกรรมศาสตรมหาบัณฑิต]. จุฬาลงกรณ์มหาวิทยาลัย, 2552

[40] A. David, and V. Sergei, “k-means++: The Advantages of Careful Seeding”. Stanford University, Supported in part by NDSEG Fellowship, NSF Grant ITR-0331640, and grants from Media-X and SNRC.

ภาคผนวก

ภาคผนวก ก

ผลการวิจัยของวิธีการที่นำเสนอ ใน 300 รอบของการทำซ้ำ

1. ผลการวิจัยของวิธีการที่นำเสนอจากข้อมูลที่สร้างขึ้นใน 300 รอบของการทำซ้ำ ประกอบไปด้วยชุดข้อมูลที่มีการแจกแจงปรกติมาตรฐาน การแจกแจงแบบเลขชี้กำลัง และชุดข้อมูลที่มีการแจกแจงเอกรูป โดยขนาดข้อมูลฝึกใน 300 รอบของทั้ง 3 การแจกจง แต่ละรอบมีขนาดข้อมูลฝึกที่แตกต่างกัน เช่น N(0,1) ขนาด (N×Feature) คือ 1,000,000×4 มีขนาดข้อมูลฝึกตั้งแต่ 92 Case ถึง 131 Case แทนด้วย Case : Case

การแจกแจงของข้อมูล	ขนาดข้อมูล (N×Feature)	พารามิเตอร์	ขนาดข้อมูลฝึก	ประสิทธิภาพความแม่น	
		สัดส่วน	Case : Case	Mean of Accuracy %	Mean of AUC
N(0,1)	1,000,000×4	0.05	92 : 131	96.7033	0.9524
	100,000×9	0.10	69 : 101	93.5488	0.9049
	30,000×29	0.10	39 : 68	89.7434	0.8319
	7,000×74	0.10	40 : 57	89.9560	0.8001
exp(1)	1,000,000×4	0.05	70 : 126	95.4507	0.9431
	100,000×9	0.10	71 : 103	87.4997	0.8579
	30,000×29	0.10	56 : 88	69.7312	0.6762
	7,000×74	0.10	46 : 83	60.718	0.6044
U(0,1)	1,000,000×4	0.05	37 : 83	93.0384	0.9072
	100,000×9	0.10	79 : 117	91.8149	0.8907
	30,000×29	0.10	60 : 93	80.6771	0.7862
	7,000×74	0.10	52 : 81	73.78905	0.715736

หมายเหตุ : กำหนดพารามิเตอร์ Km และ RT = 10 ในการหาข้อมูลฝึก

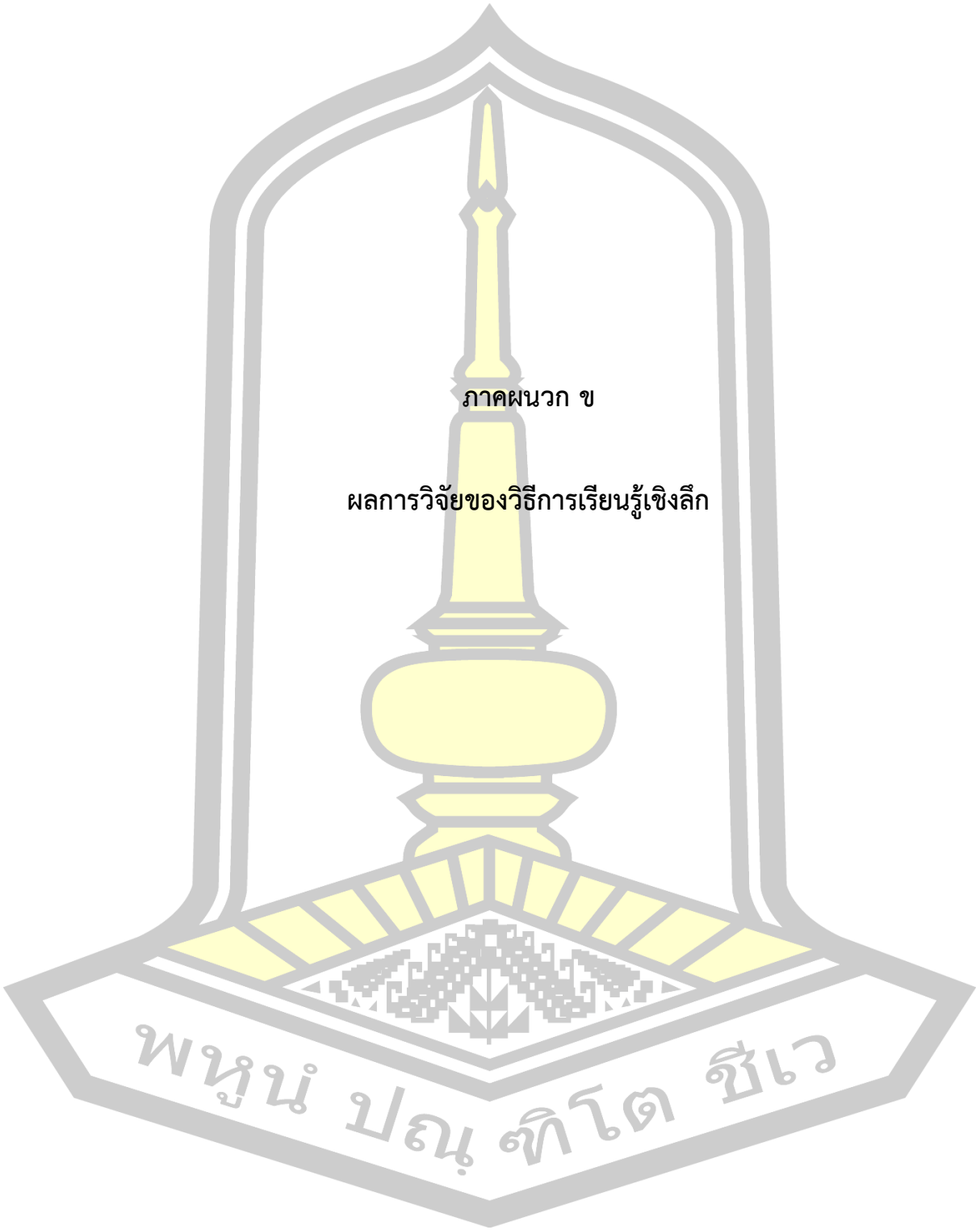
2. ผลการวิจัยของวิธีการที่นำเสนอจากชุดข้อมูลที่มีการแจกแจงแบบเลขชี้กำลัง เมื่อพารามิเตอร์เปลี่ยนไป ใน 300 รอบของการทำซ้ำ

การแจกแจงของข้อมูล	ขนาดข้อมูล (N×Feature)	พารามิเตอร์ Ratio	ขนาดข้อมูลฝึก Case : Case	ประสิทธิภาพความแม่น Mean of Accuracy %	Mean of AUC
exp(0.5)	1,000,000×4	0.05	69 : 127	95.4512	0.9443
	100,000×9	0.10	65 : 104	87.5285	0.8550
	30,000×29	0.10	56 : 87	69.4995	0.6751
	7,000×74	0.10	47 : 87	60.1873	0.6056
exp(1.5)	1,000,000×4	0.05	69 : 127	95.4502	0.9443
	100,000×9	0.10	71 : 102	87.5332	0.8566
	30,000×29	0.10	55 : 88	69.6835	0.6776
	7,000×74	0.10	52 : 83	59.7918	0.6085
exp(2)	1,000,000×4	0.05	69 : 120	95.4393	0.9470
	100,000×9	0.10	66 : 105	87.6698	0.8541
	30,000×29	0.10	55 : 91	69.6545	0.6751
	7,000×74	0.10	48 : 82	60.8629	0.6055

หมายเหตุ : กำหนดพารามิเตอร์ Km และพารามิเตอร์ RT = 10 ในการหาข้อมูลฝึก

ภาคผนวก ข

ผลการวิจัยของวิธีการเรียนรู้เชิงลึก



1. ผลการวิจัยของวิธีการเรียนรู้เชิงลึกที่ใช้ข้อมูลฝึกขนาดเท่ากับวิธีการที่ศึกษา จากข้อมูลที่สร้างขึ้นใน 300 รอบของการทำซ้ำ ประกอบไปด้วยชุดข้อมูลที่มีการแจกแจงปรกติมาตรฐาน การแจกแจงแบบเลขชี้กำลัง และชุดข้อมูลที่มีการแจกแจงเอกรูป โดยขนาดข้อมูลฝึกใน 300 รอบของทั้ง 3 การแจกจง แต่ละรอบมีขนาดข้อมูลฝึกที่แตกต่างกัน เช่น N(0,1) ขนาด (N×Feature) คือ 1,000,000×4 มีขนาดข้อมูลฝึกตั้งแต่ 92 Case ถึง 131 Case แทนด้วย Case : Case

การแจกแจงของข้อมูล	ขนาดข้อมูล (N×Feature)	พารามิเตอร์	ขนาดข้อมูลฝึก	ประสิทธิภาพความแม่น	
		Ratio	Case : Case	Mean of Accuracy %	Mean of AUC
N(0,1)	1,000,000×4	0.05	92 : 131	96.4298	0.9479
	100,000×9	0.10	69 : 101	93.1439	0.8958
	30,000×29	0.10	39 : 68	89.7429	0.8366
	7,000×74	0.10	40 : 57	89.1348	0.8208
exp(1)	1,000,000×4	0.05	70 : 126	95.1148	0.9414
	100,000×9	0.10	71 : 103	87.6143	0.8514
	30,000×29	0.10	56 : 88	70.6632	0.6787
	7,000×74	0.10	46 : 83	61.6226	0.6091
U(0,1)	1,000,000×4	0.05	37 : 83	91.9004	0.9086
	100,000×9	0.10	79 : 117	91.3858	0.8874
	30,000×29	0.10	60 : 93	80.8719	0.7863
	7,000×74	0.10	52 : 81	73.3911	0.7293

ประวัติผู้เขียน

ชื่อ	นางสาวนันทัชพร เสนาวงค์
วันเกิด	วันที่ 4 ธันวาคม พ.ศ. 2538
สถานที่เกิด	อำเภอโพนทอง จังหวัดร้อยเอ็ด
สถานที่อยู่ปัจจุบัน	บ้านเลขที่ 10 หมู่ 9 ตำบลโคกกกม่วง อำเภอโพนทอง จังหวัดร้อยเอ็ด รหัสไปรษณี 45110
ประวัติการศึกษา	พ.ศ. 2561 ปริญญาวิทยาศาสตรบัณฑิต (สถิติ) มหาวิทยาลัยมหาสารคาม
ผลงานวิจัย	การเปรียบเทียบรูปแบบการพยากรณ์ปริมาณน้ำฝนรายเดือนในภาคตะวันออกเฉียงเหนือ