การเปรียบเทียบประสิทธิภาพวิธีการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์

EFFICIENCY COMPARISONS OF MACHINE LEARNING AND DEEP LEARNING METHODS IN SPAM MAIL DETECTION

จิรายุ พิทักษ์ตันสกุล

ดิษย์ฉัตร เปียลาวัณ

ปีย์วรานุช เบญญาบุณาพจน์

ปัญหาพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรบัณฑิต (สถิติประยุกต์)

ภาควิชาสถิติ คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ปีการศึกษา 2565

EFFICIENCY COMPARISONS OF MACHINE LEARNING AND DEEP LEARNING METHODS IN SPAM MAIL DETECTION

JIRAYU PHITHUKTUNSAKUL
DISACHAT PIALAWAN
PEEWARANUCH BENYABUNAPOJ

A SPECIAL PROBLEM SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENT FOR
THE DEGREE OF BACHELOR OF SCIENCE (APPLIED STATISTICS)
DEPARTMENT OF STATISTICS, SCHOOL OF SCIENCE
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
ACADEMIC YEAR 2022

หัวข้อปัญหาพิเศษ	การเปรียบเทียบประสิทธิภาพวิธีการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์ Efficiency Comparison of Machine Learning and Deep Learning Methods in Spam Mail Detection		
ชื่อนักศึกษา	นายจิรายุ พิทักษ์ตันสกุล	รหัสนักศึกษา	62050758
	นายดิษย์ฉัตร เปียลาวัณ	รหัสนักศึกษา	62050772
	นางสาวปีย์วรานุช เบญญาบุณาพจน์	รหัสนักศึกษา	62050801
ปริญญา	วิทยาศาสตรบัณฑิต (สถิติประยุกต์)		
ภาควิชา	สถิติประยุกต์		
ปีการศึกษา	2565		
อาจารย์ที่ปรึกษา	รศ.สายชล สินสมบูรณ์ทอง		

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

คณะกรรมการสอบ	ลายมือชื่อ
ผศ.พรชัย หลายพสุ ประธานกรรมการ	
ผศ.ดร.ยุวดี กล่อมวิเศษ กรรมการ	
รศ.สายชล สินสมบูรณ์ทอง ประธานกรรมการ	

ลิขสิทธิ์ของคณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

หัวข้อปัญหาพิเศษ การเปรียบเทียบประสิทธิภาพวิธีการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์

ชื่อนักศึกษา นายจิรายุ พิทักษ์ตันสกุล รหัสนักศึกษา 62050758
นายดิษย์ฉัตร เปียลาวัณ รหัสนักศึกษา 62050772
นางสาวปีย์วรานุช เบญญาบุณาพจน์ รหัสนักศึกษา 62050801

ปริญญา วิทยาศาสตรบัณฑิต (สถิติประยุกต์)

ภาควิชา สถิติประยุกต์

คณะ วิทยาศาสตร์

มหาวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง (สจล.)

ปีการศึกษา 2565

อาจารย์ที่ปรึกษา รศ.สายชล สินสมบูรณ์ทอง

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาวิธีการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก สำหรับประยุกต์ใช้ในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์ โดยข้อมูลที่นำมาศึกษาได้นำมาจากเว็บไซต์ kaggle กระบวนการวิเคราะห์ข้อมูลเริ่มจากการจัดเตรียมข้อมูลจากนั้นทำการนำข้อมูลที่จัดเตรียมแล้วไปกำจัดค่านอกเกณฑ์ 3 วิธี คือ วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF แล้วทำการแบ่งข้อมูลเป็นชุดข้อมูลฝึกสอน 70% และชุดข้อมูลทดสอบ 30% ใช้แบบจำลองการเรียนรู้ของเครื่อง 3 วิธี คือ วิธีการเรียนรู้นาอีฟเบส์ วิธีซัพพอร์ตเวกเตอร์แมชชีน และวิธีเพื่อนบ้านใกล้สุด k ตัว ใช้แบบจำลองการเรียนรู้เชิงลึก 3 วิธี คือ วิธีโครงข่ายประสาทเทียม วิธีโครงข่ายแบบวนซ้ำ และวิธีหน่วยความจำระยะสั้นยาว แล้วนำเข้าการเรียนรู้แบบรวมกลุ่มด้วยวิธีการโหวตทั้งการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก ผลการวิจัยพบว่าการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN ให้ค่าประสิทธิภาพการทำนายดีสุด โดยพิจารณาจากค่าความแม่น ค่าความเที่ยง ค่าเรียกคืน และค่าประสิทธิภาพโดยรวมที่สูงสุด วิธีการเรียนรู้ของเครื่องโดยใช้การเรียนรู้แบบรวมกลุ่มให้ประสิทธิภาพการทำนายดังนี้ 98.13% 95.46% 98.22% และ 96.82% ตามลำดับ วิธีการเรียนรู้เชิงลึกโดยใช้การเรียนรู้แบบรวมกลุ่มให้ ดังนี้ 98.26% 96.89% 97.11% และ 97.00% ตามลำดับ สรุปได้ว่าการเรียนรู้เชิงลึกให้ประสิทธิภาพการทำนายได้ดีกว่าแบบจำลองการเรียนรู้ของเครื่อง

คำสำคัญ : จดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์, ค่านอกเกณฑ์, การเรียนรู้ของเครื่อง, การเรียนรู้เชิงลึก, การเรียนรู้แบบรวมกลุ่ม

Title	EFFICIENCY COMPARISONS OF MACHINE LEARNING AND DEEP LEARNING METHODS IN SPAM MAIL DETECTION
Students	Mr. Jirayu Phithuktunsakul Student ID 62050758
	Mr. Disachat Pialawan Student ID 62050772
	Miss Peewaranuch Benyabunapoj Student ID 62050801
Degree	Bachelor of Science (Applied Statistics)
Department	Statistics
School	Science
University	King Mongkut's Institute of Technology Ladkrabang (KMITL)
Academic Year	2022
Advisor	Assoc.Prof. Saichon Sinsomboonthong

Abstract

The purpose of this research paper is to study machine learning and deep learning techniques for detecting spam emails. The dataset used in the study was obtained from the Kaggle and applied three outlier removal methods, namely DBSCAN, Isolation Forest, and IF-LOF, after data preparation. The data was split into training and test sets at a 70:30 ratio before applying three machine learning methods, namely Naive Bayes, Support Vector Machine, and k-Nearest Neighbors, and three deep learning methods, namely Artificial Neural Network, Recurrent Neural Network, and Long Short-Term Memory. Ensemble learning was also utilized to combine machine learning and deep learning methods. Results of the study revealed that the DBSCAN outlier removal method demonstrated the highest predictive performance with accuracy, precision, recall, and F-measure scores of 98.13%, 95.46%, 98.22%, and 96.82% for machine learning methods, and 98.26%, 96.89%, 97.11%, and 97.00% for deep learning methods, respectively. In summary, deep learning provides better predictive performance than machine learning models.

Keywords : Spam mail, Outlier, Machine learning, Deep learning, Ensemble Learning

กิตติกรรมประกาศ

ปัญหาพิเศษฉบับนี้สำเร็จลุล่วงไปได้ด้วยดีเนื่องด้วยได้รับความกรุณาจาก รศ.สายชล สินสมบูรณ์ทอง อาจารย์ที่ปรึกษาปัญหาพิเศษที่กรุณาสละเวลาอันมีค่าให้คำแนะนำ ข้อเสนอแนะ คำปรึกษาและช่วยปรับปรุงแก้ไขข้อบกพร่องต่างๆด้วยความเอาใจใส่อย่างดียิ่ง อีกทั้งแนะแนวความรู้ต่างๆเอื้อเฟื้อเอกสารอ้างอิงในการค้นคว้าข้อมูลและติดตามความก้าวหน้าของงานทุกขั้นตอน จนทำให้ปัญหาพิเศษฉบับนี้เสร็จสมบูรณ์ คณะผู้วิจัยตระหนักถึงความตั้งใจ ความเมตตาและความทุ่มเทของอาจารย์ จึงขอกราบขอบพระคุณด้วยความเคารพอย่างสูงไว้ ณ โอกาสนี้

ขอขอบพระคุณ ผศ.พรชัย หลายพสุ และผศ.ดร.ยุวดี กล่อมวิเศษ ที่ให้เกียรติเป็นคณะกรรมการปัญหาพิเศษที่กรุณาให้คำปรึกษา แนวคิดข้อเสนอแนะที่เป็นประโยชน์ต่อการจัดทำปัญหาพิเศษฉบับนี้ และสละเวลาตรวจทานชี้ให้เป็นถึงข้อบกพร่องต่างๆทำให้ปัญหาพิเศษฉบับนี้มีความสมบูรณ์มากยิ่งขึ้น

รวมถึงขอกราบขอบพระคุณ คณาจารย์ประจำภาควิชาสถิติประยุกต์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง (สจล.) ที่ได้ให้ความรู้ ความเข้าใจอีกทั้งคำแนะนำและความช่วยเหลือการประสานงานต่างๆ อย่างสม่ำเสมอ

สุดท้ายนี้ขอกราบขอบพระคุณ บิดา มารดา ที่สนับสนุนและส่งเสริมกำลังใจเสมอมาและขอขอบคุณเพื่อนๆ ที่ให้คำปรึกษาและช่วยเหลือตลอดจนทำให้ปัญหาพิเศษฉบับนี้สำเร็จตามที่ได้ตั้งใจ และคณะผู้วิจัยหวังอย่างยิ่งว่าปัญหาพิเศษฉบับนี้จะเป็นประโยชน์และเป็นแนวทางต่อไป

จิรายุ พิทักษ์ตันสกุล
ดิษย์ฉัตร เปียลาวัณ
ปีย์วรานุช เบญญาบุณาพจน์

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญรูป

รูปที่ **หน้า**

สารบัญรูป (ต่อ)

รูปที่ **หน้า**

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญ

สังคมในปัจจุบันมีการเปลี่ยนแปลงไปอย่างรวดเร็ว โดยเฉพาะด้านการส่งจดหมายอิเล็กทรอนิกส์ (Mail) เพื่อแลกเปลี่ยนข้อความของผู้ส่งและผู้รับ มีผู้ที่ใช้งานจดหมายอิเล็กทรอนิกส์อยู่มากกว่า 4 พันล้านคนทั่วโลก เนื่องจากทุกวันนี้ผู้คนมีโทรศัพท์ที่สามารถส่งจดหมายอิเล็กทรอนิกส์ได้ มีสัญญาณอินเตอร์เน็ตตามสถานที่ต่างๆ เช่น ห้างสรรพสินค้า สวนสาธารณะ และสถานที่ทำงาน ซึ่งจดหมายอิเล็กทรอนิกส์สามารถใช้งานได้ง่าย รวดเร็ว และมีค่าใช้จ่ายน้อย และยังสามารถแนบรูปหรือเอกสารข้อมูลต่างๆส่งไปได้ จึงเป็นที่นิยมใช้ในการติดต่อสื่อสารเป็นอย่างมาก ทั้งทางด้านธุรกรรม และการโฆษณาต่างๆ อีกทั้งยังมีการใช้บัญชีของจดหมายอิเล็กทรอนิกส์ในการลงทะเบียนเพื่อเข้าใช้บริการทางอินเทอร์เน็ตอีกมากมาย

จึงอาจเป็นช่องโหว่ในการส่งจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์ (Spam mail) อันเป็นจดหมายอิเล็กทรอนิกส์ที่มักเกิดจากการได้ข้อมูลส่วนบุคคลของผู้อื่น แล้วมีการนำข้อมูลเหล่านั้นไปใช้ในการติดต่อ อาจทั้งเพื่อการโฆษณาประชาสัมพันธ์ขายสินค้าหรือการบริการ และยังเพื่อเชิญชวนเข้าร่วมกิจกรรมต่างๆ และในบางกรณีอาจมีการใช้เพื่อเผยแพร่ข้อมูลหรือภาพอันไม่เหมาะสม ตลอดจนอาจจะมีการนำข้อมูลที่ได้มานั้นไปใช้กระทำความผิดในรูปแบบต่างๆ (กฤษฎา, 2556) ซึ่งจะทำให้ผู้รับสูญเสียข้อมูลส่วนตัวได้ ซึ่งจะเกิดความเสียหายต่อข้อมูลส่วนตัวและยังต้องเสียเวลาไปในการไปกู้คืนข้อมูลที่เสียไปมา หรืออาจเกิดความเสียหายต่อข้อมูลในคอมพิวเตอร์และอุปกรณ์คอมพิวเตอร์ เนื่องด้วยการพัฒนาการทางเทคโนโลยีและระบบเครือข่ายอินเตอร์เน็ตมีความเจริญก้าวหน้าไปอย่างรวดเร็ว การส่งจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์อาจทำได้โดยการเขียนโปรแกรมสั้นๆ ที่ทำการคัดลอกจดหมายอิเล็กทรอนิกส์ต้นฉบับเป็นจำนวนมากๆ หรือเป็นการแพร่กระจายไวรัสไปสู่อุปกรณ์ที่ใช้ในการส่งจดหมายอิเล็กทรอนิกส์ ทำให้สูญเสียข้อมูลส่วนตัวไป เนื่องจากการให้บริการจดหมายอิเล็กทรอนิกส์เป็นแบบกระจายศูนย์ (Decentralized) จึงเป็นการยากที่จะควบคุมการส่งจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์ต่างจากการใช้งานโทรศัพท์หรือการส่งจดหมายทั่วไป ซึ่งเป็นการบริการแบบรวมศูนย์ (Centralized) คือผู้ใช้บริการทั้ง 2 แบบ ต้องทำการติดต่อหรือส่งผ่านศูนย์กลางที่ใช้ในการให้บริการ จึงทำให้ง่ายต่อการควบคุมมากกว่า (วิศาล, 2549) การป้องกันจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์จึงเป็นเรื่องที่สำคัญในการป้องกันข้อมูลส่วนตัว ซึ่งข้อมูลส่วนตัวอาจทำให้เสียหายไปสู่ทรัพย์สิน อาจทำให้ต้องเสียเงินและเป็นการยากที่จะสามารถตามเงินที่เสียไปคืนมา รวมไปถึงข้อมูลเลขบัตรเครดิตหรือทรัพย์สินอื่นๆ ถ้าเกิดหลงเชื่อจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์นี้

ดังนั้นคณะผู้วิจัยสนใจที่จะศึกษาการเปรียบเทียบประสิทธิภาพวิธีการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์ โดยใช้การตรวจจับค่านอกเกณฑ์ 3 วิธี คือ วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว (Isolation Forest) และวิธี IF-LOF (Isolation Forest- Local Outlier Factor) ใช้การเรียนรู้ของเครื่อง 3 วิธี คือ วิธีนาอีฟเบส์ (Naive Bayes) วิธีซัพพอร์ตเวกเตอร์ แมชชีน (Support Vector Machine) และวิธีเพื่อนบ้านใกล้สุด k ตัว (K-Nearest Neighbors) และใช้การเรียนรู้เชิงลึก 3 วิธี คือ วิธีโครงข่ายประสาทเทียม (Artificial Neural Network) วิธีโครงข่ายประสาทแบบวนซ้ำ (Recurrent Neural Network) และวิธีหน่วยความจำระยะสั้นยาว (Long Short-Term Memory Networks) แล้วใช้การเรียนรู้แบบรวมกลุ่ม (Ensemble Learning) ทั้งวิธีการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก ซึ่งวิธีที่กล่าวมาข้างต้นเป็นวิธีที่มีประสิทธิภาพที่ดี โดยคณะผู้วิจัยจะพิจารณาค่าความแม่น (Accuracy) ค่าความเที่ยง (Precision) ค่าเรียกคืน (Recall) และค่าประสิทธิภาพโดยรวม (F-Measure) เพื่อนำมาเปรียบเทียบว่าวิธีไหนให้ประสิทธิภาพดีที่สุด

1.2 วัตถุประสงค์

1) ศึกษาวิธีการตรวจจับค่านอกเกณฑ์ในประเภทของข้อมูลที่เป็นข้อความ
2) ศึกษาวิธีการเรียนรู้ของเครื่องและวิธีการเรียนรู้เชิงลึกในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์
3) เปรียบเทียบวิธีการเรียนรู้ของเครื่องและวิธีการเรียนรู้เชิงลึก
4) เปรียบเทียบวิธีการเรียนรู้แบบรวมกลุ่มในวิธีการเรียนรู้ของเครื่องและวิธีการเรียนรู้เชิงลึก

1.3 ขอบเขตของงานวิจัย

1) ขอบเขตด้านข้อมูล

   นำข้อมูลมาจากเว็บไซต์ kaggle.com โดยข้อมูลมีตัวแปรที่ใช้ทั้งหมด 2 ตัวแปรคือ ข้อความ (text) และหมายเลขคำตอบของข้อมูล (label_num) ทั้งหมด 5,171 ข้อความ

2) ขอบเขตด้านเนื้อหา

   2.1 วิธีตรวจจับค่านอกเกณฑ์ 3 วิธี คือ วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF

   2.2 วิธีการเรียนรู้ของเครื่อง 3 วิธี คือ วิธีนาอีฟเบส์ วิธีซัพพอร์ตเวกเตอร์แมชชีน และวิธีเพื่อนบ้านใกล้สุด k ตัว

   2.3 วิธีการเรียนรู้เชิงลึก 3 วิธี คือ วิธีโครงข่ายประสาทเทียม วิธีโครงข่ายประสาทแบบวนซ้ำ และวิธีหน่วยความจำระยะสั้นยาว

2.4 การเปรียบเทียบวิธีการเรียนรู้ของเครื่องและวิธีการเรียนรู้เชิงลึกจากการตรวจจับค่านอกเกณฑ์โดยใช้วิธีการเรียนรู้แบบรวมกลุ่ม

3) ขอบเขตด้านเครื่องมือ

3.1 โปรแกรม Microsoft excel เวอร์ชัน 365

3.2 โปรแกรม Visual Studio Code เวอร์ชัน 1.73

3.3 ภาษาที่ใช้ในการเขียนโปรแกรม คือ Python

4) ขอบเขตด้านระยะเวลา

ระยะเวลาดำเนินงานระหว่าง ตุลาคม พ.ศ. 2565 ถึง มิถุนายน พ.ศ.2566

1.4 ประโยชน์ที่คาดว่าจะได้รับ

เพื่อทราบถึงประสิทธิภาพของแต่ละแบบจำลองทั้งในการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกว่าแบบจำลองไหนมีประสิทธิภาพดีกว่ากันในการตรวจจับจดหมายอิเล็กทรอนิกไม่พึงประสงค์

1.5 นิยามศัพท์

ในงานวิจัยครั้งนี้ ผู้วิจัยได้กำหนดความหมายของคำศัพท์ที่เกี่ยวข้องไว้ เพื่อทำให้ผู้ที่จะนำงานวิจัยนี้ไปศึกษาต่อได้เกิดความเข้าใจในแนวทางเดียวกัน ดังนี้

1) จดหมายอิเล็กทรอนิกไม่พึงประสงค์ (Spam Mail) เป็นจดหมายอิเล็กทรอนิกส์ที่มักเกิดจากการได้ข้อมูลส่วนบุคคลของผู้อื่น แล้วมีการนำข้อมูลเหล่านั้นไปใช้ในการติดต่อ อาจทั้งเพื่อการโฆษณาประชาสัมพันธ์ขายสินค้าหรือบริการ และยังเพื่อเชิญชวนเข้ารวมกิจกรรมต่างๆ และในบางกรณีอาจมีการใช้เพื่อเผยแพร่ข้อมูล หรือภาพอันไม่เหมาะสม ตลอดจนอาจจะมีการนำข้อมูลที่ได้มานั้นไปใช้กระทำความผิดในรูปแบบต่างๆ (กฤษฎา, 2556)
2) การเรียนรู้ของเครื่อง (Machine Learning) คือ การทำให้ระบบคอมพิวเตอร์เรียนรู้และสร้างขั้นตอนวิธี (Algorithm) ที่สามารถเรียนรู้ข้อมูลและทำนายข้อมูลได้ (วีระพันธ์, 2564)
3) การเรียนรู้เชิงลึก (Deep Learning) เป็นหนึ่งในฟังก์ชันของปัญญาประดิษฐ์ (AI) ที่เรียนแบบการทำงานของสมองมนุษย์ในกระบวนการประมวลผลข้อมูลและเป็นการสร้างรูปแบบสำหรับใช้ในการตัดสินใจ (ณัฏฐณิชา, 2562)
4) วิธีเพื่อนบ้านใกล้สุด k ตัว (K-Nearest Neighbors) เป็นการเปรียบเทียบความคล้ายกันของข้อมูลที่สนใจกับข้อมูลอื่นว่ามีความคล้ายกันหรืออยู่ใกล้กับข้อมูลใดมากที่สุด k ตัว จากนั้นจะทำการตัดสินใจว่าคำตอบของข้อมูลที่สนใจนั้นควรเป็นคำตอบเดียวกับข้อมูลที่อยู่ใกล้ที่สุด k ตัวนั้น โดยที่ k คือความถี่ของข้อมูลที่อยู่ใกล้กับข้อมูลที่สนใจ (พัชนา, 2561)
5) วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) คือ การหาค่าความถูกต้องในการจำแนกข้อมูลภาพและการกระจายตัวของข้อมูลต้องอยู่บนข้อสมมติ (Assumption) ของ

การกระจายแบบจำลองปกติที่ให้ค่าความถูกต้องของการจำแนกข้อมูลภาพมีความถูกต้องแม่นยำสูง (Mountrakis et al., 2011)

6) วิธีนาอีฟเบส์ (Naive Bayes) คือ การเรียนรู้ของเครื่องที่อาศัยหลักการความน่าจะเป็นตามทฤษฎีเบส์ (Bayes Theorem) ซึ่งมีขั้นตอนวิธี (Algorithm) ที่ไม่ซับซ้อน เป็นขั้นตอนวิธีในการจำแนกข้อมูลโดยการเรียนรู้ปัญหาที่เกิดขึ้นเพื่อนำมาสร้างเงื่อนไขการจำแนกข้อมูลใหม่ (อนัตต์ชัย และจรัญ, 2561)
7) วิธีโครงข่ายประสาทเทียม (Artificial Neural Network) เป็นเทคนิคที่เลียนแบบการทำงานของสมองมนุษย์ซึ่งประกอบด้วยเซลล์ประสาท (Neuron) และแต่ละเซลล์จะถูกเชื่อมโยงกันเป็นโครงข่ายซึ่งในซอฟต์แวร์ เซลล์ประสาทจะเรียกว่าโหนด (Node) และแต่ละโหนดจะถูกแบ่งออกเป็นชั้น (Layer) โดยหลักการของการเรียนรู้เชิงลึกก็จะเป็นวิธีโครงข่ายประสาทเทียมที่มีโหนดหลายๆ ชั้น ทำให้สามารถประมวลผลได้ครั้งละจำนวนมาก ช่วยให้การเรียนรู้ของเครื่องสามารถให้ผลลัพธ์ในการตัดสินใจและคาดการณ์ได้ดีมากยิ่งขึ้น (ณัฏฐณิชา, 2562)
8) วิธีโครงข่ายประสาทแบบวนซ้ำ (Recurrent Neural Network) มีหลักการทำงานคือการนำข้อมูลออกหรือผลลัพธ์ (Output Data) ที่ได้จากการคำนวณจากโหนดก่อนหน้านี้กลับมาใช้เป็นข้อมูลเข้า (Input Data) ของโหนดถัดไป ซึ่งแต่ละโหนดของวิธีโครงข่ายแบบวนซ้ำนั้นจะมีข้อมูลที่เข้ามา 2 ส่วน คือ ข้อมูลเข้าของโหนดนั้นๆ กับข้อมูลออกที่ผ่านการคำนวณจากโหนดก่อนหน้า โดยข้อมูลทั้ง 2 ชุดที่เข้ามาในโหนดจะถูกรวมเข้าด้วยกัน ก่อนจะถูกแยกผลลัพธ์ออกเป็น 2 ส่วน คือ ผลลัพธ์ที่ได้จากโหนดนั้น ๆ และผลลัพธ์ที่จะถูกนำไปเป็นข้อมูลเข้าของโหนดถัดไป วิธีโครงข่ายแบบวนซ้ำนั้นเหมาะนำมาใช้งานกับข้อมูลที่มีลักษณะเป็นลำดับ (Sequence) หรือข้อมูลที่มีความต่อเนื่อง (วิศรุต และวริศ, 2563)
9) วิธีหน่วยความจำระยะสั้นยาว (Long Short-Term Memory) เป็นโครงข่ายแบบวนซ้ำรูปแบบหนึ่งที่ถูกพัฒนาขึ้นมาให้มีความเสถียรและมีประสิทธิภาพมากขึ้น โดยหลักการทำงานคือสามารถเก็บสถานะหรือข้อมูลของแต่ละโหนดเอาไว้เพื่อที่เวลาย้อนกลับไปพิจารณา จะได้ทราบถึงที่มาของข้อมูลค่าดังกล่าว (วิศรุต และวริศ, 2563)
10) ค่านอกเกณฑ์ (Outlier) เป็นจุดข้อมูลที่มีการกระจายตัวแตกต่างจากจุดข้อมูลปกติในชุดข้อมูล (Hossain et al., 2021)

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

การวิจัยครั้งนี้ได้ทำการศึกษาการเปรียบเทียบประสิทธิภาพวิธีการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์ โดยมีการตรวจจับค่านอกเกณฑ์ 3 วิธี คือ วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว (Isolation Forest) และวิธี IF-LOF (Isolation Forest-Local Outlier Factor) ก่อนนำไปใช้การเรียนรู้ของเครื่อง 3 วิธี คือ วิธีนาอีฟเบส์ (Naive Bayes) วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) และวิธีเพื่อนบ้านใกล้สุด k ตัว (K-Nearest Neighbors) แล้วใช้การเรียนรู้เชิงลึก 3 วิธี คือวิธีโครงข่ายประสาทเทียม (Artificial Neural Network) วิธีโครงข่ายประสาทแบบวนซ้ำ (Recurrent Neural Network) และวิธีหน่วยความจำระยะสั้นยาว (Long Short-Term Memory Networks) ซึ่งใช้ค่าความแม่น (Accuracy) ค่าความเที่ยง (Precision) ค่าเรียกคืน (Recall) และค่าประสิทธิภาพโดยรวม (F-Measure) ในการเปรียบเทียบประสิทธิภาพของแบบจำลอง แล้วใช้การเรียนรู้แบบรวมกลุ่มทั้งวิธีการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกในการเพิ่มประสิทธิภาพ

2.1 ทฤษฎีที่ใช้ในการจัดเตรียมข้อมูล

2.1.1 การเปลี่ยนตัวอักษรพิมพ์ใหญ่เป็นตัวอักษรพิมพ์เล็ก

เป็นการลดขนาดของคำโดยการเปลี่ยนให้ตัวอักษรทั้งหมดเป็นตัวอักษรพิมพ์เล็ก เนื่องจากคอมพิวเตอร์นั้นมีการเก็บค่าของตัวอักษรพิมพ์ใหญ่และตัวอักษรพิมพ์เล็กที่แตกต่างกัน เช่น รหัสมาตรฐานของสหรัฐอเมริกาเพื่อการแลกเปลี่ยนสารสนเทศ (American Standard Code for Information Interchange : ASCII) ตัวอักษร “A” มีค่าเป็น “65” และตัวอักษร “a” มีค่าเป็น “97” เป็นต้น ตัวอักษรพิมพ์ใหญ่มักจะถูกใช้ในการบ่งบอกถึงคำนามที่มีความเฉพาะเจาะจง และยังถูกใช้เป็นคำย่อที่เกิดจากการนำตัวอักษรตัวแรกหรือกลุ่มอักษรบางตัวมาสร้างเป็นคำใหม่ ในขณะที่ตัวอักษรพิมพ์เล็กข้อมูลไม่ได้มีแนวโน้มทางความหมายใดๆ การแปลงตัวอักษรให้เป็นตัวอักษรพิมพ์เล็กจึงเป็นประโยชน์เพราะทำให้มิติของข้อมูลมีขนาดเล็กลง เพิ่มประสิทธิภาพทางสถิติ และไม่ลดความสมบูรณ์ของข้อมูลไป (Hickman et al., 2022)

2.1.2 การลบคำฟุ่มเฟือย (Stop Word)

คำฟุ่มเฟือยเป็นส่วนหนึ่งของภาษาธรรมชาติ คำฟุ่มเฟือยเป็นคำที่ไม่มีนัยสำคัญมักปรากฏขึ้นบ่อยในข้อความ ซึ่งคำฟุ่มเฟือยไม่มีประโยชน์ต่อการจำแนกข้อความจึงสมควรลบออกไป ทำให้สามารถลดเวลาในการประมวลผลของคอมพิวเตอร์ได้ โดยประเภทของคำฟุ่มเฟือยในภาษาอังกฤษ

ได้แก่ คำบุพบท คำสรรพนาม เป็นต้น ยกตัวอย่างเช่น “the”, “in”, “a”, “an” (Mohan et al., 2015)

2.1.3 การลบตัวอักขระพิเศษ

การลบตัวอักขระพิเศษรวมไปถึงสัญลักษณ์ วรรคตอนและตัวเลข ช่วยให้สามารถให้ความสนใจกับคำหรือวลีในข้อความได้มากขึ้น ซึ่งเครื่องหมายวรรคตอนสามารถคงไว้ได้ ในขณะที่ตัวเลขและอักขระพิเศษถูกลบออก (Hickman et al., 2022) ตัวอย่างอักขระพิเศษ เช่น ! - # /

2.1.4 Word2vec

เป็นแบบจำลองที่ใช้สร้างการฝังคำหรือแปลงคำให้อยู่ในรูปแบบของเวกเตอร์ ซึ่งเวกเตอร์ของคำต่างๆ ถูกคำนวณจากบริบทรอบข้าง โดยใช้เทคนิคโครงข่ายประสาทเทียม (Neural Network) แบบ Encoder-Decoder มีชั้น (Layer) จำนวน 2 ชั้น ซึ่งมีหลักการในการเปรียบเทียบเวกเตอร์ทางความหมายของคำทั้ง 2 คำ แล้วคืนค่าออกมาเป็นตัวเลขตั้งแต่ -1 ถึง 1 ซึ่งบ่งชี้ถึงความใกล้เคียงทางความหมายโดยให้ค่าจากน้อยไปมาก พูดอีกนัยหนึ่งว่าคำที่มีบริบทการปรากฏคล้ายๆ กันควรเป็นคำที่มีความหมายคล้ายกันด้วย ซึ่งวิธี Word2vec สามารถวัดค่าความคล้ายทางความหมายของเวกเตอร์ของคำและใช้ร่วมกับการจำแนกข้อมูลประเภทข้อความได้ดี (บุษบงก์ และคณะ, 2564)

2.1.5 การลดมิติเวกเตอร์ของคำโดยใช้การวิเคราะห์ส่วนประกอบหลัก (Principal Component Analysis)

เป็นการหารูปแบบความสัมพันธ์ระหว่างตัวแปรของข้อมูลเพื่อนำไปทำกระบวนการลดมิติของข้อมูลโดยที่ยังไม่ทำให้สูญเสียข้อมูลสำคัญไป การวิเคราะห์ส่วนประกอบหลักจัดเป็นการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Machine Learning Algorithm) เพราะไม่ได้มีการนำเอาผลเฉลยของข้อมูล (Label) มาพิจารณาร่วมด้วย ซึ่งเป็นเทคนิคหนึ่งที่ใช้กระบวนการทางสถิติและเมทริกซ์ (Matrix) เพื่อเข้ามาช่วยอธิบายข้อมูลให้เป็นที่เข้าใจได้ง่ายยิ่งขึ้นเพื่อลดความซับซ้อนลง (ภัครพล, 2564)

2.2 การกำจัดค่านอกเกณฑ์

2.2.1 วิธี DBSCAN (Density-based spatial clustering of applications with noise)

เป็นวิธีการหนึ่งในการแบ่งกลุ่ม (Clustering) โดยไม่ต้องกำหนดจำนวนกลุ่ม (Cluster) ที่ต้องการแบ่งเหมือนกับ K-means จึงทำให้ DBSCAN เหมาะสำหรับข้อมูลที่ไม่เป็นกลุ่มก้อนหรือการกระจายตัวไม่มีรูปแบบ (No Pattern) ที่มีแกนกลาง (Core) ลักษณะเป็นรูปทรงต่างๆ ที่ K-means ไม่สามารถจัดกลุ่มได้ อีกทั้งเหมาะสำหรับการตัดค่านอกเกณฑ์ (Outlier) หรือข้อมูลรบกวน (Noise) ออกไป (สุทธิพงศ์, 2561)

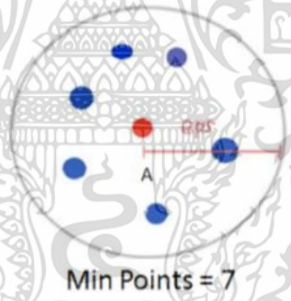
จุดแกนกลาง (Core point) คือ จุดที่มีลักษณะข้อมูลอยู่ใกล้กันเป็นจำนวนมาก ถ้าแกนกลางรวมกันได้ก็จะรวมกัน (Merge) เป็นกลุ่มเดียวกัน

DBSCAN เป็นการหาบริเวณที่ข้อมูลเกาะกลุ่มกัน ซึ่งสามารถคำนวณได้จากจุดข้อมูลที่อยู่ในบริเวณรอบ ๆ ในรัศมีที่กำหนด ซึ่งการที่จะใช้ DBSCAN ได้ จำเป็นต้องมีพารามิเตอร์ 2 ตัว คือ

1. รัศมีจากจุดศูนย์กลาง (Epsilon)
2. จำนวนจุดข้อมูลขั้นต่ำ (Min Points) สำหรับการกำหนดจุดศูนย์กลาง (Center)

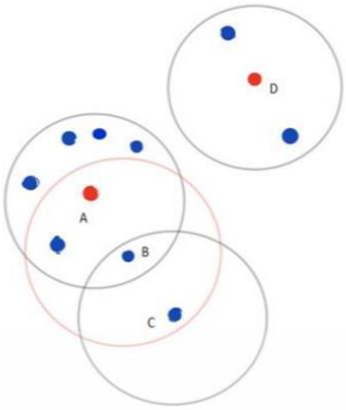
ลักษณะการทำงานของ DBSCAN

1) ในแต่ละจุดของข้อมูลจะทำการคำนวนหาจุดของข้อมูลที่อยู่ใกล้กันทั้งหมดในรัศมีจากจุดศูนย์กลาง (Epsilon) ถ้าจุดข้อมูลไหนมีจุดข้อมูลที่อยู่ใกล้กันมากกว่าหรือเท่ากับจำนวนจุดข้อมูลขั้นต่ำ (Min Points) จะให้จุดนั้นเป็นจุดแกนกลาง (Core point) และสร้างเป็นกลุ่มใหม่ (Cluster)
2) ในแต่ละจุดแกนกลาง ถ้ามีจุดข้อมูลที่อยู่ใกล้กันกับที่เชื่อมต่อกับอีกจุดแกนกลางได้ให้รวมเป็นกลุ่มใหม่
3) ถ้าจุดข้อมูลใดไม่เชื่อมต่อกับจุดแกนกลางก็จะให้จุดข้อมูลนั้นเป็นค่านอกเกณฑ์ (Outlier) ซึ่งไม่อยู่ในกลุ่มใดๆ



รูปที่ 2.1 การกำหนดจุดแกนกลางสำหรับการทำ DBSCAN โดยกำหนดให้จำนวนจุดข้อมูลขั้นต่ำเท่ากับ 7

จากรูปที่ 2.1 ตำแหน่ง A จะเรียกว่าจุดแกนกลางเพราะในรัศมีจาก A นั้นมีจุดข้อมูลที่อยู่ใกล้กันอย่างน้อย 6 จุด ถ้านำจุดแกนกลางไปรวมกับจุดข้อมูลที่อยู่ในรัศมี A ก็จะได้ Min Point เท่ากับ 7



รูปที่ 2.2 การกำหนดจำนวนการจัดกลุ่มข้อมูลในการทำ DBSCAN

จากรูปที่ 2.2 จะแสดงลักษณะในการกำหนดตำแหน่งต่างๆ เพื่อใช้ในการจัดกลุ่ม ดังนี้

ตำแหน่ง A คือ จุดแกนกลางเพราะมีจุดข้อมูลที่อยู่ใกล้กันอย่างน้อย 7

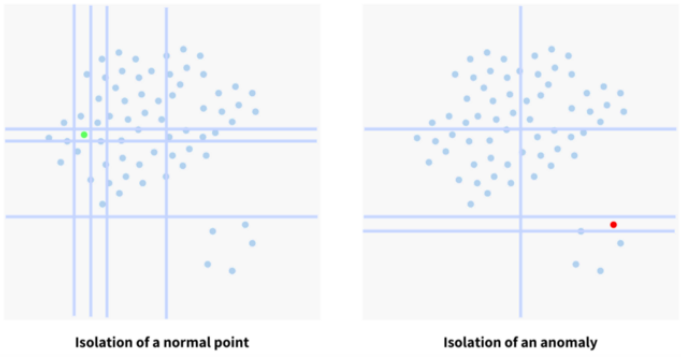
ตำแหน่ง B คือ Border เพราะ B มีจุดข้อมูลที่อยู่ใกล้กันไม่ถึง 7 แต่อยู่ในรัศมีจุดแกนกลาง A

ตำแหน่ง C คือ Border เพราะ C มีจุดข้อมูลที่อยู่ใกล้กันไม่ถึง 7 แต่ยังอยู่ในรัศมีของ B ซึ่ง B ก็ยังอยู่ในรัศมีของจุดแกนกลางของ A จึงทำให้ C ถือว่ายังอยู่ในกลุ่มเดียวกันกับ A และ B

ตำแหน่ง D คือ ค่านอกเกณฑ์ (Outlier) หรือข้อมูลรบกวน (Noise) เพราะจุดนั้นไม่ได้อยู่ในรัศมีของจุดแกนกลางใดๆเลย ซึ่งค่านอกเกณฑ์หรือข้อมูลรบกวนนั้นจะเป็นข้อมูลที่จะต้องทำการตัดออกไปและไม่รวมอยู่ในกลุ่ม (สุทธิพงศ์, 2561)

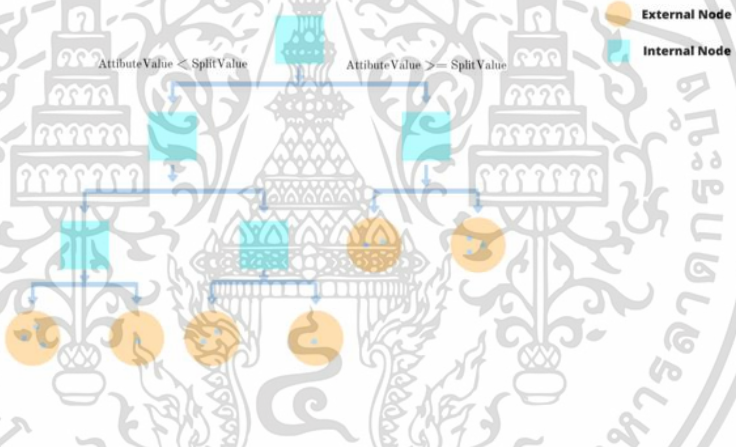
2.2.2 วิธีป่าไม้โดดเดี่ยว (Isolation Forest)

เป็นวิธีการตรวจจับค่านอกเกณฑ์ที่มีรากฐานมาจากวิธีต้นไม้ตัดสินใจ (Decision Tree) โดยเริ่มต้นจากการสุ่มคุณลักษณะ (Attribute) และแบ่งข้อมูล (Partition) ระหว่างค่าต่ำสุดและค่าสูงสุดเพื่อแยกตัวอย่าง โดยจะแบ่งข้อมูลไปเรื่อยๆ จนกระทั่งข้อมูลแต่ละตัวจะแยกจากกันโดยสมบูรณ์ วิธีป่าไม้โดดเดี่ยวถูกสร้างขึ้นจากการเพิ่มขึ้นของจำนวนต้นไม้โดดเดี่ยว (Isolation Tree) ที่ถูกแยกด้วยคุณลักษณะต่างๆ ที่แตกต่างกัน (Farzad and Gulliver, 2020)



รูปที่ 2.3 การแบ่งข้อมูลปกติ (ซ้าย) การแบ่งข้อมูลผิดปกติ (ขวา)

เมื่อข้อมูลจุดสีเทากระจายตัวในลักษณะดังรูปที่ 2.3 หากต้องการจะแบ่งข้อมูลออกจากกันสามารถทำได้โดยสร้างเส้นแบ่งขึ้นมาอาจจะเป็นแนวตั้งหรือแนวนอนก็ได้ แต่ต้องแบ่งจนกว่าจุดที่สนใจจะถูกแยกออกจากจุดอื่นโดยสิ้นเชิง ซึ่งพิจารณารูปทางซ้ายจะเห็นได้ว่าข้อมูลปกติจะอยู่กระจุกตัวกับข้อมูลจุดอื่นๆ ทำให้ต้องใช้จำนวนเส้นในการแบ่งค่อนข้างมาก เปรียบเทียบกับข้อมูลผิดปกติรูปทางขวาที่จะอยู่แยกจากข้อมูลส่วนใหญ่ ทำให้ใช้จำนวนเส้นแบ่งที่น้อยกว่าก็สามารถแยกออกมาได้



รูปที่ 2.4 ต้นไม้ตัดสินใจ (Decision Tree)

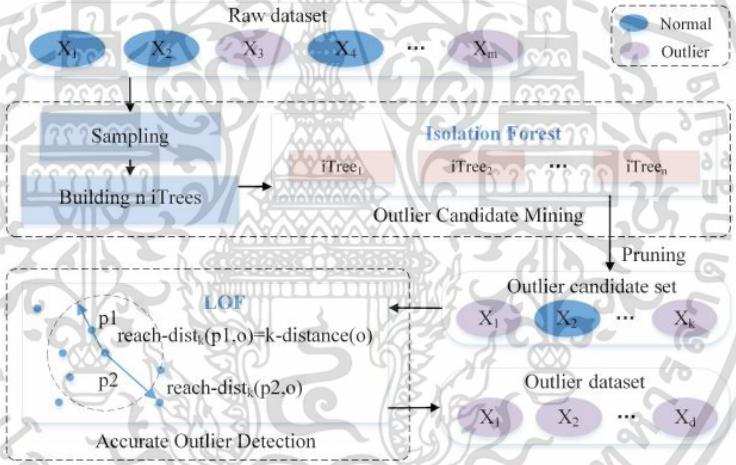
จากรูปที่ 2.4 จะเห็นได้ว่าเส้นแบ่งแต่ละเส้นก็คือเส้นที่ตัดแบ่งข้อมูลทั้งหมดออกเป็นกิ่งซ้ายและกิ่งขวา แล้วทำการแบ่งไปจนกระทั่งถึงจุดที่ทุกอย่างแยกออกจากกัน ทำให้ได้ต้นไม้ที่มีกิ่งแยกข้อมูลออกจากกัน หลังจากนั้นจะพบว่าข้อมูลปกติจะใช้จำนวนชั้นของต้นไม้ที่ลึกมากในการแบ่งข้อมูลให้เป็นอิสระจากกัน แต่ข้อมูลผิดปกติจะถูกกรองตั้งแต่ชั้นแรกๆ ของต้นไม้ ทำให้ความลึกของข้อมูลที่ผิดปกติจะตื้นกว่าข้อมูลอื่นๆ

เมื่อทราบความลึกของต้นไม้แล้วทำให้สามารถคำนวณเป็นคะแนนความผิดปกติ (Anomaly score) เพื่อใช้ในการแยกประเภทของข้อมูลได้ ซึ่งคะแนนความความผิดปกตินั้นจะมีค่าตั้งแต่ 0 ถึง 1 โดยค่าที่เข้าใกล้ 1 จะหมายถึงข้อมูลที่มีแนวโน้มผิดปกติ ส่วนข้อมูลที่มีค่าน้อยกว่า 0.5 ลงไปจะถือว่าเป็นข้อมูลทั่วไปที่ไม่มีความผิดปกติ (อาณัตชัย, 2564)

2.2.3 วิธี IF-LOF (Isolation Forest- Local Outlier Factor)

เป็นวิธีแก้ไขปัญหาการตรวจจับค่านอกเกณฑ์ที่อ่อนไหวต่อความผิดปกติที่มีความหลากหลาย (Global Outlier) และใช้เวลาในการประมวลผลนาน โดยจะทำการตัด (Prune) ชุดข้อมูลแทนการใช้ชุดข้อมูลดั้งเดิมและนำมาใช้เป็นแหล่งข้อมูล (Data Source) จึงสามารถลดปริมาณข้อมูลที่ต้องประมวลผลได้เป็นอย่างมาก โดยมีวิธีการดังนี้ (Zhangyu et al., 2019)

1) นำข้อมูลดิบไปใช้ในวิธีป่าไม้โดดเดี่ยวซึ่งถูกสร้างขึ้นจากการเพิ่มขึ้นของจำนวนต้นไม้โดดเดี่ยว (Isolation Tree) ที่ถูกแยกด้วยคุณลักษณะต่างๆ ที่แตกต่างกัน จากนั้นคำนวณคะแนนความผิดปกติจากความลึกของต้นไม้
2) ทำการตัดกิ่ง (Pruning) โดยตัดจุดข้อมูลปกติบางส่วนออกตามเกณฑ์การตัดแต่งกิ่งเพื่อให้ได้ชุดข้อมูลผิดปกติที่เหลืออยู่
3) คำนวณค่า LOF แต่ละจุดข้อมูลที่มาจากชุดข้อมูลผิดปกติที่เหลืออยู่และเลือก d จุดแรกที่มีค่า LOF สูงเป็นค่านอกเกณฑ์



รูปที่ 2.5 แผนภาพการทำงานของวิธี IF-LOF

2.3 การเรียนรู้ของเครื่อง (Machine Learning)

2.3.1 วิธีนาอีฟเบส์ (Naive Bayes)

เป็นวิธีที่มีขั้นตอนวิธีการทำงานที่ไม่ซับซ้อน โดยจะใช้หลักการของความน่าจะเป็น (Probability) ซึ่งมีพื้นฐานมาจากทฤษฎีเบส์ (Bayes Theorem) หรือทฤษฎีที่ว่าด้วยโอกาสที่จะเกิดขึ้นของเหตุการณ์ต่างๆ ซึ่งจะคำนวณความน่าจะเป็นแบบมีเงื่อนไข (Conditional Probability) (Dietrich et al., 2015) แสดงได้ดังสมการที่ 2.1

$$P(h \mid D) = \frac{P(D \mid h) \times P(h)}{P(D)} \tag{2.1}$$

D แทนข้อมูลที่นำมาใช้ในการคำนวณการแจกแจงความน่าจะเป็นภายหลัง (Posteriori Probability) ของการเกิดเหตุการณ์ h คือ $P(h|D)$

โดยที่ $P(h)$ คือ ค่าความน่าจะเป็นก่อน (Prior probability) ของการเกิดเหตุการณ์ h

$P(D)$ คือ ค่าความน่าจะเป็นก่อนของชุดข้อมูลตัวอย่าง D

$P(h|D)$ คือ ค่าความน่าจะเป็นของ h เมื่อรู้ D

$P(D|h)$ คือ ค่าความน่าจะเป็นของ D เมื่อรู้ h

กำหนดให้ $P(h)$ คือความน่าจะเป็นที่จะเกิดเหตุการณ์ h และ $P(h|D)$ คือความน่าจะเป็นที่จะเกิดเหตุการณ์ h เมื่อเกิดเหตุการณ์ D แล้วจากตัวแปรที่กำหนด และแนวคิดของเบส์นั้นเราสามารถทำนายเหตุการณ์ที่พิจารณาได้จากการเกิดของเหตุการณ์ต่างๆ ได้ดังสมการที่ 4 ข้างต้น

2.3.2 วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

เป็นวิธีหนึ่งที่ได้รับความนิยมอย่างมากในงานที่เกี่ยวกับการจดจำรูปแบบ ตลอดจนการแก้ปัญหาการจำแนกกลุ่ม (Classification Problem) โดยใช้การหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งกลุ่มของข้อมูลที่เข้า และมาผ่านกระบวนการสอนให้ระบบเรียนรู้ โดยเน้นไปยังเส้นแบ่งแยกกลุ่มข้อมูลได้ดีที่สุด (Optimal Separating Hyperplane) เมื่อเราพิจารณาข้อมูลที่มี 2 กลุ่ม (Wang et al., 2009) ดังสมการที่ 2.2

$$D = \{(x_i, y_i); i = 1, 2, ..., n\} \tag{2.2}$$

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

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

ซึ่งเป็นการกำหนดกลุ่มเป้าหมายให้ซัพพอร์ตเวกเตอร์แมชชีนโดยที่ซัพพอร์ตเวกเตอร์แมชชีนมุ่งเป้าหมายเพื่อหาฟังก์ชันการตัดสินใจที่สามารถแบ่งแยกค่าที่ไม่ทราบได้ ดังสมการที่ 2.3

$$f(x) = sign\{\sum_{k=1}^{n_v} w_k \varphi_k(x)\varphi_k(x_k) + b\} \tag{2.3}$$

เมื่อ $$\varphi(x) = [\varphi_1(x_1), \varphi_2(x_2), ..., \varphi_n(x_{n_v})]^T \tag{2.4}$$

กลุ่มข้อมูล x จากสมการที่ 2.2 ซึ่งไม่สามารถแบ่งแยกโดยใช้สมการเส้นตรงได้ แต่จะถูกแปลงให้อยู่ในรูปแบบที่สามารถใช้สมการเส้นตรงในการแบ่งได้ โดยใช้ฟังก์ชันเคอร์เนล (Kernel Function) ดังสมการที่ 2.5

$$K(x, x_k) = \varphi(x)\varphi(x_k) \tag{2.5}$$

เมื่อ $\varphi(x)$ แทน ฟังก์ชันสำหรับแปลงข้อมูลที่ไม่เป็นเชิงเส้นให้เป็นข้อมูลที่อยู่ในรูปเชิง

เส้นสามารถแบ่งแยกได้

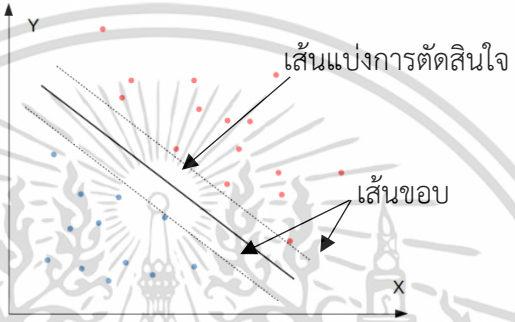
w_k แทน ค่าน้ำหนักที่เชื่อมโยงจาก feature space ไปสู่ output space

b แทน ค่าเอนเอียง (Bias)

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

n_v แทน จำนวนซัพพอร์ตเวกเตอร์

การเพิ่มเส้นขอบ (Margin) เป็นวิธีที่ดีที่สุดในการหาเส้นแบ่งของข้อมูล ซึ่งการสร้างเส้นขอบที่สัมผัสกับค่าของข้อมูลใน Feature Space ที่ใกล้ที่สุด ดังนั้นยิ่งเส้นขอบมีขนาดกว้างขึ้น เส้นแบ่งจะถือว่าเป็นเส้นที่ดี และเรียกตำแหน่งการสัมผัสข้อมูลที่ใกล้ที่สุดจากการเพิ่มขอบนี้ว่าซัพพอร์ตเวกเตอร์ (Support Vector)



รูปที่ 2.6 การใช้ซัพพอร์ตเวกเตอร์แมชชีนในการจำแนกข้อมูลสองกลุ่ม (Binary classification)

จากรูปที่ 2.6 เป็นการจำแนกข้อมูลออกเป็น 2 กลุ่มคือสีน้ำเงินและสีแดง โดยวิธีซัพพอร์ตเวกเตอร์แมชชีนจะทำการหาเส้นแบ่งการตัดสินใจที่เป็นเส้นทึบ ซึ่งเส้นนี้จะเกิดขึ้นระหว่างกลางของเส้นประทั้งด้านซ้ายและขวา โดยมีเงื่อนไขว่าจะต้องหาคู่ของเส้นประที่กว้างที่สุดเท่าที่จะเป็นไปได้ เนื่องจากในบางกรณีการแบ่งแยกกลุ่มไม่สามารถทำได้ถูกต้องโดยสมบูรณ์ ดังนั้นจึงต้องมีการกำหนดตัวแปรสำหรับยอมรับค่าความผิดพลาดโดยการเพิ่มตัวแปร ξ (Slack variable) ดังสมการที่ 2.6 และ 2.7 ดังนี้

$$w^T x + b \geq y - \xi_i \quad \text{เมื่อกำหนดให้ } y = 1 \tag{2.6}$$

$$w^T x + b \leq y + \xi_i \quad \text{เมื่อกำหนดให้ } y = -1 \tag{2.7}$$

จากการกำหนดค่า $\xi_i > 0$ ทำให้โครงสร้างของซัพพอร์ตเวกเตอร์แมชชีนบรรลุวัตถุประสงค์ใน 2 ส่วน คือ การเพิ่มระยะแบ่งแยกให้มากที่สุดและลดข้อผิดพลาดในการทำนายให้ต่ำที่สุด ดังสมการที่ 2.8

$$Minimize \frac{1}{2}\|W\|^2 + c\sum_{i=1}^{N} \xi_i \tag{2.8}$$

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

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

ฟังก์ชันเคอร์เนลที่นิยมใช้มีอยู่ 3 ชนิดด้วยกันคือ

1. ฟังก์ชันโพลิโนเมียล (Polynomial Function) ไว้ใช้สำหรับต้องการแบ่งเส้นข้อมูล 2 ฝั่งที่ไม่เป็นกลุ่ม มีเส้นโค้งตรงกลาง เหมาะใช้กับข้อมูลที่มีเส้นที่ซับซ้อน

$$K(x_i, x_j) = (x_i^T x_j + r)^\gamma ; \gamma > 0 \quad (2.9)$$

2. ฟังก์ชันเกาส์เซียนเรเดียลเบสิส (Gaussian Radial Basis Function-RBF) ไว้ใช้สำหรับข้อมูลที่เป็นกลุ่ม

$$K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2); \gamma > 0 \quad (2.10)$$

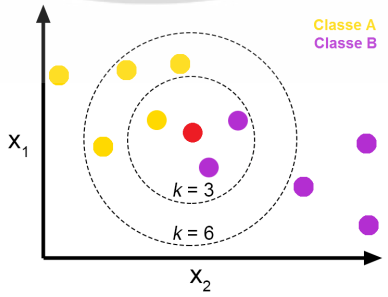
3. ฟังก์ชันซิกมอยด์ (Sigmoid Function) ไว้สำหรับต้องการแบ่งข้อมูล 2 ฝั่ง เหมาะกับข้อมูลที่มีลักษณะกระจายออกจากกัน

$$K(x_i, x_j) = \tanh(\gamma x_i^T x_j - r); \gamma > 0 \quad (2.11)$$

โดยฟังก์ชันเคอร์เนลที่นำมาใช้คือ ฟังก์ชันเกาส์เซียนเรเดียลเบสิส (Gaussian Radial Basis Function-RBF) คือ การเรียนรู้ปรับค่าน้ำหนักให้ได้ฟังก์ชันการส่งที่เหมาะสมที่สุด ซึ่งผลตอบสนองของฟังก์ชันขึ้นอยู่กับระยะห่างระหว่างข้อมูลเข้ากับจุดศูนย์กลางของฟังก์ชัน คือ ถ้าระยะห่างใกล้จุดศูนย์กลาง ข้อมูลออกจะมาก แต่ถ้าอยู่ห่าง ข้อมูลออกที่ได้จะลดลงตามลำดับ ดังนั้น RBF จึงเหมาะในงานการประมาณค่าฟังก์ชัน ฟังชันที่นิยมใช้ใน RBF มากที่สุดคือ ฟังก์ชันเกาส์เซียน (Gaussian Function) โดยมีพารามิเตอร์การกระจาย (Spread Parameter) เป็นตัวควบคุมความกว้างของ RBF (กาญจนา, 2561)

2.3.3 วิธีเพื่อนบ้านใกล้สุด k ตัว (K-Nearest Neighbors)

จะเปรียบเทียบความคล้ายคลึงกันของข้อมูลที่ไม่รู้ว่าอยู่กลุ่มใด กับข้อมูลอื่นว่ามีความคล้ายคลึงหรืออยู่ใกล้กับข้อมูลใดมากที่สุด k ตัว จากนั้นจะทำการตัดสินว่าคำตอบของข้อมูลที่ไม่รู้ว่าอยู่กลุ่มใด นั้นควรเป็นคำตอบเดียวกับข้อมูลที่อยู่ใกล้ที่สุด k ตัวนั้น โดยที่ k คือความถี่ของข้อมูลที่อยู่ใกล้กับข้อมูลที่ไม่รู้ว่าอยู่กลุ่มใด สามารถทำได้โดยกำหนดค่า k จากนั้นคำนวณหาระยะห่างระหว่างข้อมูลตัวอย่างที่สนใจกับข้อมูลอื่นๆ ทุกตัวด้วยวิธีระยะห่างยุคลิด (Euclidian Distance) ดังสมการที่ 2.12 (พัชนา, 2561)



รูปที่ 2.7 วิธีเพื่อนบ้านใกล้สุด k ตัว โดยค่า k เท่ากับ 3 และ 6

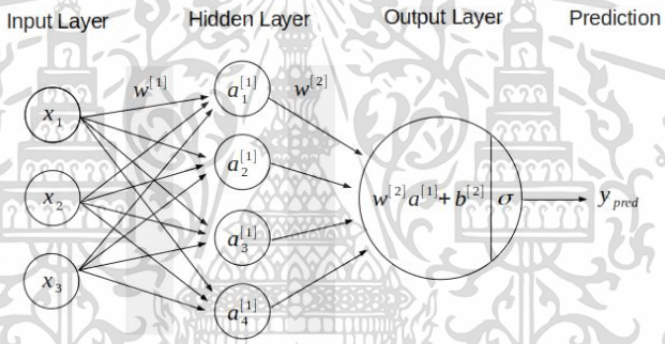
$$dist(X_i, X_j) = \sqrt{\sum_{k=1}^{n} (X_{i,k} - X_{j,k})^2} \quad (2.12)$$

โดยที่	$dist(X_i, X_j)$	คือ	ระยะห่างระหว่างตัวอย่าง X_i กับตัวอย่าง X_j
	n	คือ	จำนวนข้อมูล
	X_i	คือ	ค่าที่คำนวณได้ทั้งหมดของตัวอย่าง X_i
	$X_{i,k}$	คือ	ค่าที่คำนวณได้ตัวที่ k ของตัวอย่าง X_i
	X_j	คือ	ค่าที่คำนวณได้ทั้งหมดของตัวอย่าง X_j
	$X_{j,k}$	คือ	ค่าที่คำนวณได้ตัวที่ k ของตัวอย่าง X_j

2.4 การเรียนรู้เชิงลึก (Deep Learning)

2.4.1 วิธีโครงข่ายประสาทเทียม (Artificial Neural Network)

เป็นวิธีที่มีแนวคิดเริ่มต้นมาจากที่มีการศึกษาสมองของมนุษย์ ซึ่งประกอบด้วยจุดประสานประสาท (Synapses) และ เซลล์ประสาท (Neurons) โดยแบบจำลองนี้เกิดจากการเชื่อมต่อระหว่างเซลล์ประสาทซึ่งเรียกว่าเครือข่ายซึ่งทำงานร่วมกันได้ (ธนาภัทร, 2563)



รูปที่ 2.8 ส่วนประกอบของโครงข่ายประสาทเทียม

ชั้นข้อมูลเข้า (Input layer) มีหน้าที่รับข้อมูลเข้ามาและส่งต่อไปยังชั้นซ่อน โดยชั้นข้อมูลเข้าจะมีเพียงชั้นเดียวเท่านั้นและมีจำนวนโหนดเท่ากับข้อมูลเข้า

ชั้นซ่อน (Hidden Layer) มีหน้าที่รับข้อมูลมาจากชั้น (Layer) ก่อนหน้าและประมวณผลจึงส่งไปยังชั้นถัดไป โดยชั้นซ่อนสามารถมีได้ 1 ชั้นหรือมากกว่า ซึ่งในทางปฏิบัติเราสามารถเพิ่มหรือลดจำนวนชั้นของชั้นซ่อนเพื่อปรับความแม่นยำของแบบจำลอง

ชั้นข้อมูลออก (Output Layer) ทำหน้าที่รอรับค่าจากชั้นซ่อนโดยในชั้นข้อมูลออกจะมีจำนวนโหนดเท่ากับจำนวนค่าที่ต้องการทำนาย

2.4.2 วิธีโครงข่ายประสาทแบบวนซ้ำ (Recurrent Neural Network)

เป็นวิธีแบบโครงข่ายประสาทเทียมสำหรับการสร้างแบบจำลองที่เหมาะกับข้อมูลที่เป็นลำดับซึ่งจะมีการเก็บข้อมูลสถานะไว้ในสถานะซ่อน (Hidden State) โดยมีการนำสถานะซ่อนก่อนหน้ามาใช้ในการคำนวณสถานะซ่อนปัจจุบัน และใช้สถานะซ่อนปัจจุบันในการคำนวณข้อมูลในช่วงเวลาถัดไปโดยมีการคำนวณตามสมการที่ 2.13 และ 2.14 (เธียรศักดิ์ และธนิศา, 2561)

$$h_t = \sigma(x_t W + h_{t-1} U) \quad (2.13)$$

$$o_t = \sigma(h_t V) \quad (2.14)$$

โดยที่ t คือ ช่วงเวลา

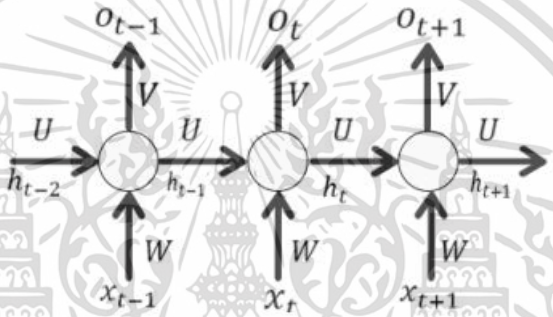
h คือ ชั้นซ่อน

x คือ ข้อมูลเข้า (Input)

o คือ ข้อมูลออก (Output)

σ คือ ฟังก์ชันกระตุ้น (Activation Function)

W, U, V คือ เมทริกซ์ถ่วงน้ำหนัก (Weight Matrix) สำหรับคำนวณข้อมูลเข้าสถานะซ่อนก่อนหน้าและสถานะซ่อนปัจจุบันตามลำดับ และมีโครงสร้างตามรูปที่ 2.9



รูปที่ 2.9 โครงสร้างของวิธีโครงข่ายประสาทแบบวนซ้ำ

แบบจำลองนี้จึงเหมาะกับข้อมูลที่มีลักษณะเป็นลำดับโดยแบบจำลองนี้สามารถใช้ได้กับข้อมูลที่มีลำดับเวลาที่ไม่ห่างกันมาก แต่จะมีปัญหาเมื่อลำดับเวลาของข้อมูลห่างกัน (Long-Term Dependencies) ซึ่งทำให้เกิดปัญหาการสูญหายของเกรเดียน (Vanishing Gradient) ตามมา

2.4.3 วิธีหน่วยความจำระยะสั้นยาว (Long Short-Term Memory)

เป็นวิธีที่แก้ปัญหาการสูญหายของเกรเดียนของวิธีโครงข่ายประสาทแบบวนซ้ำ จึงได้มีการพัฒนาวิธีหน่วยความจำระยะสั้นยาวขึ้น โดยใช้สถานะเซลล์ (Cell state) และสถานะซ่อน (Hidden state) ในการจัดเก็บข้อมูลและนำไปประมวลผลยังชั้นซ่อน ซึ่งอาศัยประตู (Gate) ต่าง ๆ ในการคำนวณว่าควรจะเก็บรักษาข้อมูลภายในสถานะเซลล์ และสถานะซ่อนมากน้อยเพียงใด โดยประกอบไปด้วยประตูข้อมูลเข้า (Input gate) ประตูข้อมูลออก (Output gate) และประตูลืม (Forget gate) ซึ่งแต่ละประตูมีหน้าที่ในการดูว่าข้อมูลที่เข้ามาควรผ่านไปหรือไม่ โดยดูจากความสำคัญของข้อมูล ซึ่งถ้ามีค่าน้อยก็ไม่สามารถผ่านไปได้ จึงช่วยป้องกันการสูญหายของเกรเดียน โดยมีการคำนวณค่าต่าง ๆ ดังสมการ (เธียรศักดิ์ และธนิศา, 2561)

$$i = \sigma(x_t W_i + h_{t-1} U_i) \quad (2.15)$$

$$f = \sigma(x_t W_f + h_{t-1} U_f) \quad (2.16)$$

$$o = \sigma(x_t W_o + h_{t-1} U_o) \quad (2.17)$$

$$c = (c_{t-1} \times f) + (i \times \sigma(x_t W_c + h_{t-1} U_c)) \quad (2.18)$$

$$h_t = \sigma(c_t) \times o \quad (2.19)$$

โดยที่ i คือ ประตูข้อมูลเข้า (Input gate)

f คือ ประตูลืม (Gorget gate)

o คือ ประตูข้อมูลออก (Output gate)

c คือ สถานะเซลล์ (Cell state)

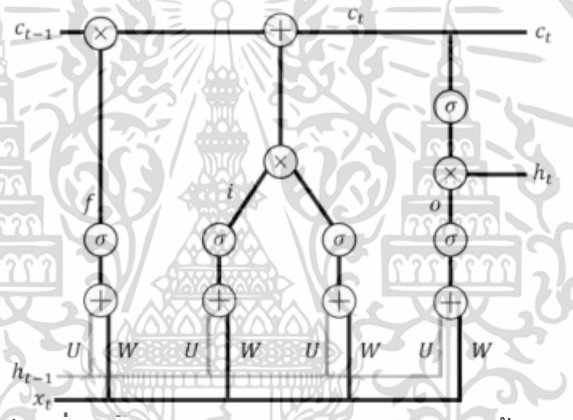
h คือ สถานะซ่อน (Hidden state)

σ คือ ฟังก์ชันกระตุ้น (Activation function)

W และ U คือ เมทริกซ์ถ่วงน้ำหนัก (Weight matrix)

t คือ ช่วงเวลา

โดยสามารถแสดงโครงสร้างได้ดังรูปที่ 2.10



รูปที่ 2.10 โครงสร้างของวิธีหน่วยความจำระยะสั้นยาว

2.5 การเรียนรู้แบบรวมกลุ่ม (Ensemble Model)

เป็นวิธีที่นำแบบจำลองในการจำแนกข้อมูล 1 ตัวหรือมากกว่า (Base Classification) ซึ่งแต่ละแบบจำลองก็จะมีกระบวนการทำงานของมันเอง และทุกตัวของแบบจำลองจะมีการจำแนกการสร้างจากกลุ่มข้อมูลเดียวกัน เมื่อได้ผลลัพธ์จากวิธีการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกในแต่ละแบบจำลองแล้ว นำผลลัพธ์ที่ได้มาผ่านวิธีการรวบรวม (Combination, Integration หรือ Vote) และตัดสินผลลัพธ์สุดท้าย (Final Decision) เพื่อให้ได้ผลลัพธ์การจำแนกข้อมูลเดียวเท่านั้น (ปรเมษฐ์ และคณะ, 2560)

การรวมแบบจำลองทำนายจะใช้วิธีการบูสต์ติง (Boosting) มีหลักการคือจะทำการสร้างแบบจำลองจำแนกกลุ่มข้อมูลหลายแบบจำลอง แต่ละแบบจำลองจะใช้ชุดข้อมูลฝึกสอนชุดเดียวกันในการสร้าง ซึ่งแต่ละแบบจำลองจะมีค่าถ่วงน้ำหนัก (Weight) เพิ่มเข้ามา โดยค่าถ่วงน้ำหนักนี้ได้มาจากความแม่น (Accuracy) ของการเรียนรู้บนชุดข้อมูลสำหรับคำตอบสุดท้ายของการทำงานด้วยวิธีบูสต์ติงจะใช้วิธีการโหวตแบบถ่วงน้ำหนักแล้วกำหนดกลุ่มให้ข้อมูลใหม่ด้วยผลโหวตที่มากที่สุด

(Majority Voting) ค่าถ่วงน้ำหนัก (Weight) เป็นค่าที่ได้จากการแปลงค่าความแม่นจากการทดสอบการทำนายของชุดข้อมูลทดสอบในแต่ละแบบจำลอง (ปรเมษฐ์ และคณะ, 2560) ดังในสมการที่ 2.20

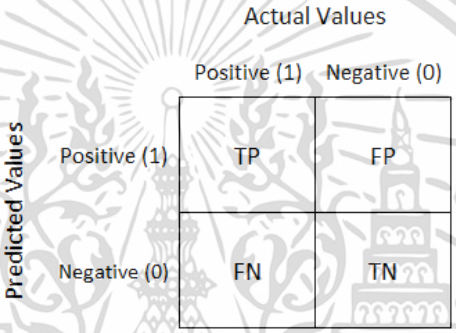
$$Weight(i) = \frac{Accuracy(i)}{100} \quad (2.20)$$

โดยที่ $Weight(i)$ คือ ค่าถ่วงน้ำหนักของแบบจำลองที่ i

$Accuracy(i)$ คือ ค่าความแม่นของแบบจำลองที่ i จากชุดข้อมูลทดสอบ

2.6 เมทริกซ์ความสับสน (Confusion Matrix)

เป็นเครื่องมือที่ใช้ในการประเมินประสิทธิภาพของแบบจำลองโดยในการทำนายจะใช้ข้อมูลที่เกิดขึ้นจริง (เอกพันธ์, 2563) ดังรูปที่ 2.11



รูปที่ 2.11 เมทริกซ์ความสับสน

บวกจริง (True Positive : TP) คือ สิ่งที่แบบจำลองทำนายว่าเกิดขึ้นจริง และเกิดขึ้นจริง

ลบจริง (True Negative : TN) คือ สิ่งที่แบบจำลองทำนายว่าเกิดขึ้นไม่จริง และเกิดขึ้นไม่จริง

บวกเท็จ (False Positive : FP) คือ สิ่งที่แบบจำลองทำนายว่าเกิดขึ้นจริง แต่เกิดขึ้นไม่จริง

ลบเท็จ (False Negative : FN) คือ สิ่งที่แบบจำลองทำนายว่าเกิดขึ้นไม่จริง แต่เกิดขึ้นจริง

1) ค่าความแม่น (Accuracy) เป็นการวัดค่าของแบบจำลองว่าทำนายถูกกี่ครั้งจากจำนวนการทำนายทั้งหมด สามารถหาได้ดังสมการที่ 2.21 (ธนาภัทร, 2563)

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (2.21)$$

2) ค่าความเที่ยง (Precision) เป็นค่าที่แบบจำลองทำนายเป็นคลาสที่สนใจ และถูกต้องต่อค่าที่แบบจำลองทำนายว่าเป็นคลาสที่สนใจทั้งถูกและผิด สามารถหาได้ดังสมการที่ 2.22 (ธนาภัทร, 2563)

$$Precision = \frac{TP}{TP + FP} \quad (2.22)$$

3) ค่าเรียกคืน (Recall) เป็นค่าที่แบบจำลองทำนายเป็นคลาสที่กำลังพิจารณาและถูกต้องต่อคลาสที่สนใจทั้งหมด สามารถหาได้ดังสมการที่ 2.23 (ธนาภัทร, 2563)

$$Recall = \frac{TP}{TP + FN} \quad (2.23)$$

4) ค่าประสิทธิภาพโดยรวม (F-Measure) เป็นการวัดความเที่ยงและค่าเรียกคืนของแบบจำลองไปพร้อม ๆ กัน สามารถหาได้ดังสมการที่ 2.24 (ธนาภัทร, 2563)

$$F-Measure = \frac{2\times precision\times recall}{precision+recall} \quad (2.24)$$

2.7 งานวิจัยที่เกี่ยวข้อง

Zhangyu et al. (2019) ได้ศึกษาการตรวจจับค่านอกเกณฑ์ด้วยวิธีป่าไม้โดดเดี่ยว (Isolation Forest) และวิธี LOF (Local Outlier Factor : LOF) เนื่องจากวิธีป่าไม้โดดเดี่ยวมีความอ่อนแอต่อการจัดการกับค่านอกเกณฑ์ในพื้นที่ (Local Outlier) ในขณะที่วิธี LOF ทำงานได้ดีในการตรวจจับค่านอกเกณฑ์ในพื้นที่ แต่ใช้เวลาในการประมวลผลนาน ผู้วิจัยจึงเสนอวิธีการเรียนรู้แบบรวมกลุ่มผสม (Two-Layer Progressive Ensemble Method) ได้แก่ วิธี IF-LOF (Isolation Forest- Local Outlier Factor) เพื่อเปรียบเทียบค่าความแม่น (Accuracy) และค่าประสิทธิภาพโดยรวม (F-Measure) กับวิธีป่าไม้โดดเดี่ยวและ วิธี LOF จากการศึกษาพบว่าวิธี IF-LOF มีค่าความแม่นและค่าประสิทธิภาพโดยรวมสูงที่สุด

Awad and Elseuofi (2011) ได้ศึกษาวิธีการเรียนรู้ของเครื่องทั้งหมด 6 วิธี ประกอบด้วย วิธีนาอีฟเบส์ (Naive Bayes) วิธีเพื่อนบ้านใกล้สุด k ตัว (K-Nearest Neighbors) วิธีโครงข่ายประสาทเทียม (Artificial Neural Network) วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) วิธีระบบภูมิคุ้มกันเทียม (Artificial Immune System) และวิธีทฤษฎีรัฟเซต (Rough Set) ในการจัดประเภทจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์ โดยวัดประสิทธิภาพจากค่าความแม่น (Accuracy) ค่าความเที่ยง (Precision) และค่าการเรียกคืน (Recall) จากการศึกษาพบว่าวิธีนาอีฟเบส์มีค่าความแม่น ค่าความเที่ยง และค่าการเรียกคืนมากที่สุด คิดเป็น 99.46%, 99.66% และ 98.64% ตามลำดับ

Poomka et al. (2019) ได้ทำการศึกษาการตรวจจับข้อความที่ไม่พึงประสงค์โดยการเรียนรู้เชิงลึกด้วยวิธีหน่วยความจำระยะสั้นยาว (Long Short-Term Memory Networks) และวิธีหน่วยเวียนกลับแบบมีประตู (Gated Recurrent Unit) โดยใช้ชุดข้อมูลจาก Almeida และ Hidalgo ที่มีข้อความทั้งหมด 5,574 ข้อความ ซึ่งประกอบด้วยข้อความที่ไม่พึงประสงค์ 747 ข้อความ และข้อความปกติ 4,827 ข้อความ โดยวัดประสิทธิภาพจากค่าความแม่น จากการศึกษาพบว่าวิธีหน่วยความจำระยะสั้นยาวมีค่าความแม่นมากที่สุดคิดเป็น 98.18%

Julis and Alajesan (2020) ได้ศึกษาการตรวจจับข้อความที่ไม่พึงประสงค์โดยใช้การเรียนรู้ของเครื่องผ่านการทำเหมืองข้อความ ใช้การเรียนรู้ของเครื่อง 5 วิธี ได้แก่ วิธีการถดถอยลอจิสติก (Logistic Regression) วิธีเพื่อนบ้านใกล้สุด k ตัว (K-Nearest Neighbors) วิธีนาอีฟเบส์ (Naive Bayes) วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) และวิธีต้นไม้ตัดสินใจ (Decision Tree) โดยวัดประสิทธิภาพจากค่าความแม่นและเวลาที่ใช้ จากการเปรียบเทียบผลลัพธ์พบว่าวิธีซัพ

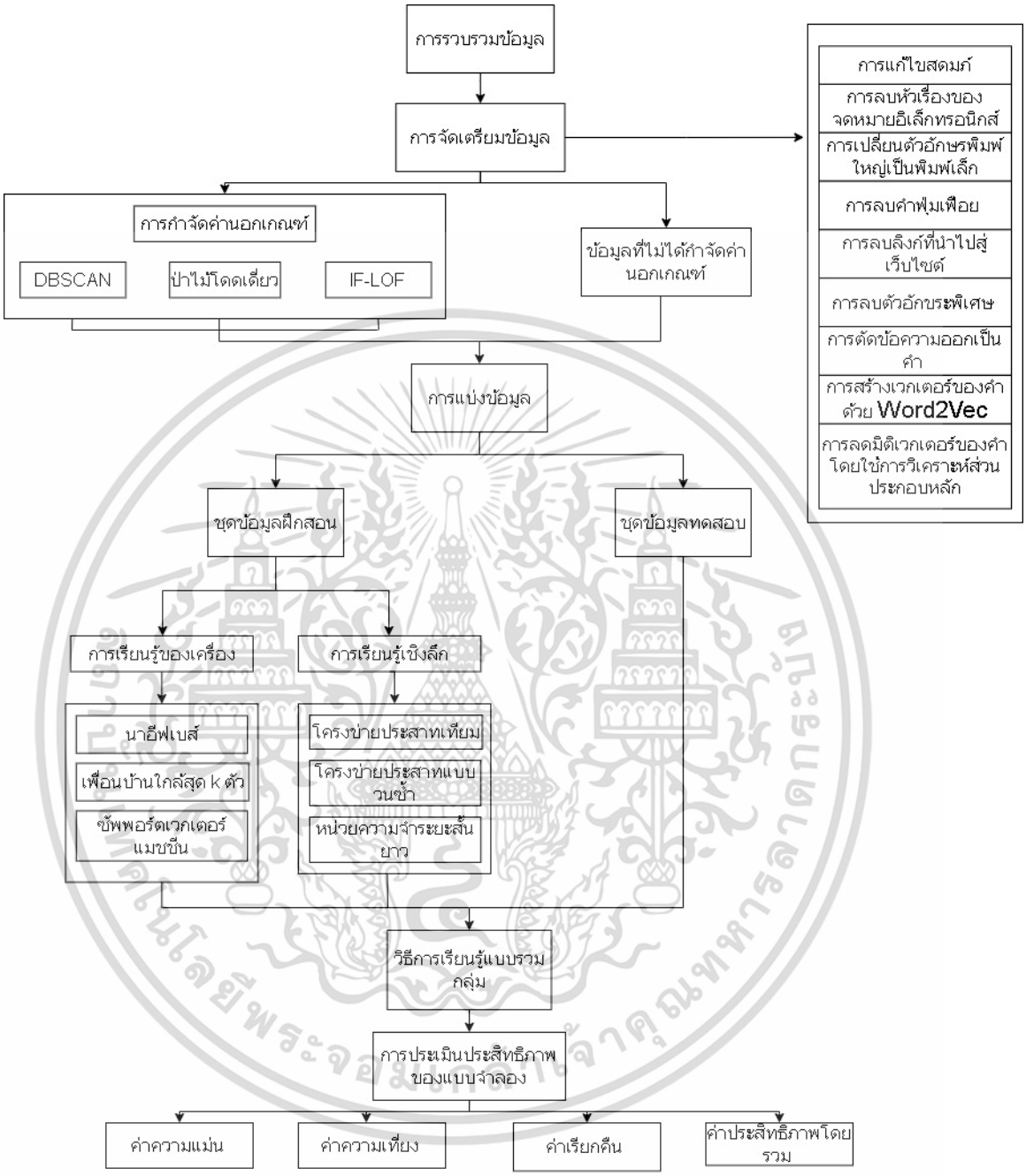
พอร์ตเวกเตอร์ แมชชีนมีค่าความแม่นมากที่สุด 98% แต่ใช้เวลาการทำนายมากที่สุด ในขณะที่วิธีนาอีฟเบส์ใช้เวลาการทำนายน้อยที่สุด แต่มีค่าความแม่นปานกลาง คิดเป็น 95%

Hossain et al. (2021) ได้ศึกษาการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์ ใช้วิธีตรวจจับค่านอกเกณฑ์ด้วยวิธีป่าไม้โดดเดี่ยวและวิธี DBSCAN ใช้การเรียนรู้ของเครื่อง 3 วิธี คือ วิธีนาอีฟเบส์อเนกนาม วิธีป่าสุ่ม และวิธีเพื่อนบ้านใกล้สุด k ตัว ใช้การเรียนรู้เชิงลึก 3 วิธี คือ วิธีโครงข่ายประสาทแบบวนซ้ำ วิธี Gradient Descent และวิธีโครงข่ายประสาทเทียม และใช้การเรียนรู้แบบรวมกลุ่มกับวิธีการเรียนรู้ของเครื่อง พบว่า วิธีการเรียนรู้ของเครื่องในการตรวจจับค่านอกเกณฑ์ด้วยวิธี DBSCAN มีค่าความแม่นเท่ากับวิธีป่าไม้โดดเดี่ยวแต่ใช้เวลาน้อยกว่า ส่วนวิธีการเรียนรู้เชิงลึกในการตรวจจับค่านอกเกณฑ์ด้วยวิธีป่าไม้โดดเดียวให้ผลดีกว่าโดยมีค่าความแม่นเท่ากับ 99.28% 99.28% และ 97.95% ตามลำดับ ซึ่งมีค่ามากกว่าการใช้วิธีการตรวจจับค่านอกเกณฑ์ด้วยวิธี DBSCAN ซึ่งมีค่าความแม่นเท่ากับ 92.42% 89.42% และ 91.42% ตามลำดับ จึงสรุปได้ว่าในการเรียนรู้ของเครื่องโดยใช้วิธีการตรวจจับค่านอกเกณฑ์ด้วยวิธี DBSCAN มีประสิทธิภาพมากที่สุด ส่วนในการเรียนรู้เชิงลึกโดยใช้วิธีการตรวจจับค่านอกเกณฑ์ด้วยวิธีป่าไม้โดดเดี่ยวมีประสิทธิภาพมากที่สุด

บทที่ 3
วิธีการดำเนินงานวิจัย

การวิจัยครั้งนี้ได้ทำการศึกษาการเปรียบเทียบประสิทธิภาพวิธีการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์ โดยมีวิธีการตรวจจับค่านอกเกณฑ์ทั้งหมด 3 วิธี คือ วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว (Isolation Forest) และวิธี IF-LOF (Isolation Forest- Local Outlier Factor) ก่อนนำไปใช้การเรียนรู้ของเครื่อง 3 วิธี คือ วิธีนาอีฟเบส์ (Naive Bayes) วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) และวิธีเพื่อนบ้านใกล้สุด k ตัว (K-Nearest Neighbors) แล้วใช้การเรียนรู้เชิงลึก 3 วิธี คือวิธีโครงข่ายประสาทเทียม (Artificial Neural Network) และวิธีโครงข่ายแบบวนซ้ำ (Recurrent Neural Network) และวิธีหน่วยความจำระยะสั้นยาว (Long Short-Term Memory Networks) ซึ่งใช้ค่าความแม่น (Accuracy) ค่าความเที่ยง (Precision) ค่าเรียกคืน (Recall) และค่าประสิทธิภาพโดยรวม (F-Measure) ในการเปรียบเทียบประสิทธิภาพของแบบจำลองแล้วใช้การเรียนรู้แบบรวมกลุ่มทั้งวิธีการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกในการเพิ่มประสิทธิภาพ โดยรายละเอียดจะถูกกล่าวถึงในหัวข้อย่อยๆตามลำดับต่อไป

3.1 ขั้นตอนการดำเนินงาน



รูปที่ 3.1 กระบวนการทำงานการเปรียบเทียบประสิทธิภาพวิธีการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก ในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์

จากรูปที่ 3.1 แสดงขั้นตอนในการดำเนินงานโดยเริ่มจากนำชุดข้อมูลมาจากเว็บไซต์ kaggle.com จากนั้นทำการจัดเตรียมข้อมูลโดยเปลี่ยนตัวอักษรข้อความพิมพ์ใหญ่เป็นพิมพ์เล็ก ลบตัวอักขระพิเศษ เช่น ! - # / และลบคำที่เป็นคำฟุ่มเฟือย (Stop Word) โดยใช้คลังคำศัพท์ภาษาอังกฤษของ nltk ขั้นตอนต่อไปทำการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF และนำไปแบ่งข้อมูลเป็น 2 ชุด คือชุดข้อมูลฝึกสอนและชุดข้อมูลทดสอบ นำชุดข้อมูลฝึกสอนมาสร้างแบบจำลองการเรียนรู้ของเครื่อง 3 วิธี คือ วิธีนาอีฟเบส์ วิธีซัพพอร์ตเวกเตอร์แมชชีน และวิธีเพื่อนบ้านใกล้สุด k ตัว การเรียนรู้เชิงลึก 3 วิธี คือวิธีโครงข่ายประสาทเทียม วิธีโครงข่ายแบบวนซ้ำ และวิธีหน่วยความจำระยะสั้นยาว แล้วนำชุดข้อมูลทดสอบมาวัดประสิทธิภาพของแบบจำลอง โดยได้ใช้วิธีการเรียนรู้แบบรวมกลุ่มทั้งการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกเพื่อเพิ่มประสิทธิภาพของแบบจำลองจากนั้นจึงนำค่าความแม่น ค่าความเที่ยง ค่าเรียกคืน และค่าประสิทธิภาพโดยรวมที่ได้มาเปรียบเทียบกันระหว่างการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก

3.2 การรวบรวมข้อมูล

นำชุดข้อมูลมาจากเว็บไซต์ kaggle.com ชุดข้อมูลชื่อ Spam Mails Dataset โดยมีข้อมูลทั้งหมด 5,171 ข้อความ ข้อมูลชุดนี้ประกอบไปด้วยจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์จำนวน 1,499 ข้อความ คิดเป็น 29% และจดหมายอิเล็กทรอนิกส์ปกติ 3,672 ข้อความ คิดเป็น 71% ซึ่งข้อมูลชุดนี้มีลักษณะดังนี้

ตารางที่ 3.1 ลักษณะของข้อมูล

ชื่อตัวแปร	คำอธิบาย	ชนิดของตัวแปร
text (x)	ข้อความในจดหมายอิเล็กทรอนิกส์	ตัวแปรเชิงคุณภาพ
label (y)	ประเภทของจดหมายอิเล็กทรอนิกส์ 0 = ham (จดหมายอิเล็กทรอนิกส์ปกติ) 1 = spam (จดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์)	ตัวแปรเชิงคุณภาพ

จากตารางที่ 3.1 ลักษณะของข้อมูลที่นำมาจากเว็บไซต์ kaggle.com จะมีตัวแปรอยู่ 2 ตัวแปร คือ text และ label ซึ่งเป็นตัวแปรที่นำมาใช้วิเคราะห์

	A	B	C	D
1	No.	label	text	label_num
2	605	ham	Subject: enron methanol ; meter # :	0
3	2349	ham	Subject: hpl nom for january 9 , 2001	0
4	3624	ham	Subject: neon retreat	0
5	4685	spam	Subject: photoshop , windows , office .	1
6	2030	ham	Subject: re : indian springs	0
7	2949	ham	Subject: ehronline web address change	0
8	2793	ham	Subject: spring savings certificate - take 30	0
9	4185	spam	Subject: looking for medication ? we ` re	1
10	2641	ham	Subject: noms / actual flow for 2 / 26	0
11	1870	ham	Subject: nominations for oct . 21 - 23 ,	0
12	4922	spam	Subject: vocable % rnd - word asceticism	1

รูปที่ 3.2 ตัวอย่างชุดข้อมูลจดหมายอิเล็กทรอนิกส์

จากรูปที่ 3.2 ข้อมูลที่นำมาจากเว็บไซต์ kaggle.com จะมีตัวแปรอยู่ 4 ตัวแปรคือ No, label, text และ label_num ซึ่งตัวแปรที่นำมาวิเคราะห์มี 2 ตัวแปรคือ text และ label_num

3.3 การจัดเตรียมข้อมูล

3.3.1 การนำข้อมูลเข้า

ทำการนำชุดข้อมูลจดหมายอิเล็กทรอนิกส์จากเว็บไซต์ kaggle.com เข้าโปรแกรม Visual Studio Code โดยใช้ชุดคำสั่ง pandas ในการอ่านไฟล์นามสกุล csv.

3.3.2 การแก้ไขสดมภ์

3.3.2.1 การลบสดมภ์

ใช้คำสั่ง drop จากชุดคำสั่ง pandas ในการลบสดมภ์ Unnamed:0 และ label เนื่องจากเป็นตัวแปรที่ไม่ได้ใช้ในการวิจัย ตัวอย่างชุดข้อมูลจดหมายอิเล็กทรอนิกส์หลังจากลบสดมภ์ Unnamed:0 และ label แสดงดังรูปที่ 3.3 และ 3.4

	Unnamed: 0	label	text	label_num
0	605	ham	Subject: enron methanol ; meter # : 988291\r\n...	0
1	2349	ham	Subject: hpl nom for january 9 , 2001\r\n(see...	0
2	3624	ham	Subject: neon retreat\r\nho ho ho , we ' re ar...	0
3	4685	spam	Subject: photoshop , windows , office . cheap ...	1
4	2030	ham	Subject: re : indian springs\r\nthis deal is t...	0
...	...	...	...	...
5166	1518	ham	Subject: put the 10 on the ft\r\nthe transport...	0
5167	404	ham	Subject: 3 / 4 / 2000 and following noms\r\nhp...	0
5168	2933	ham	Subject: calpine daily gas nomination\r\n>\r\n...	0
5169	1409	ham	Subject: industrial worksheets for august 2000...	0
5170	4807	spam	Subject: important online banking alert\r\ndea...	1

รูปที่ 3.3 ตัวอย่างชุดข้อมูลจดหมายอิเล็กทรอนิกส์ก่อนลบสดมภ์ Unnamed:0 และ label

จากรูปที่ 3.3 ทำการลบสดมภ์ของตัวแปร Unnamed:0 และ label เนื่องจากไม่ได้นำมาใช้วิเคราะห์ และลดขนาดของข้อมูลที่นำมาวิเคราะห์

	text	label_num
0	Subject: enron methanol ; meter # : 988291\r\n...	0
1	Subject: hpl nom for january 9 , 2001\r\n(see...	0
2	Subject: neon retreat\r\nho ho ho , we ' re ar...	0
3	Subject: photoshop , windows , office . cheap ...	1
4	Subject: re : indian springs\r\nthis deal is t...	0
...	...	...
5166	Subject: put the 10 on the ft\r\nthe transport...	0
5167	Subject: 3 / 4 / 2000 and following noms\r\nhp...	0
5168	Subject: calpine daily gas nomination\r\n>\r\n...	0
5169	Subject: industrial worksheets for august 2000...	0
5170	Subject: important online banking alert\r\ndea...	1

รูปที่ 3.4 ตัวอย่างชุดข้อมูลจดหมายอิเล็กทรอนิกส์หลังจากลบสดมภ์ Unnamed:0 และ label

จากรูปที่ 3.4 ทำการลบตัวแปรให้เหลือ 2 ตัวแปร คือ ตัวแปร text เป็นตัวแปรของข้อความในจดหมายอิเล็กทรอนิกส์ และ ตัวแปร label_num เป็นตัวแปรที่บอกว่าเป็นจดหมายอิเล็กทรอนิกส์ปกติหรือจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์

3.3.2.2 การเปลี่ยนชื่อสดมภ์

ใช้คำสั่ง columns จากชุดคำสั่ง pandas ในการเปลี่ยนชื่อตัวแปร label_num เป็น label เพื่อความเข้าใจง่ายและชัดเจนในการสร้างแบบจำลอง ตัวอย่างชุดข้อมูลจดหมายอิเล็กทรอนิกส์หลังจากเปลี่ยนชื่อสดมภ์ label_num เป็น label แสดงดังรูปที่ 3.5 และ 3.6

	text	label_num
0	Subject: enron methanol ; meter # : 988291\r\n...	0
1	Subject: hpl nom for january 9 , 2001\r\n(see...	0
2	Subject: neon retreat\r\nho ho ho , we ' re ar...	0
3	Subject: photoshop , windows , office . cheap ...	1
4	Subject: re : indian springs\r\nthis deal is t...	0
...	...	...
5166	Subject: put the 10 on the ft\r\nthe transport...	0
5167	Subject: 3 / 4 / 2000 and following noms\r\nhp...	0
5168	Subject: calpine daily gas nomination\r\n>\r\n...	0
5169	Subject: industrial worksheets for august 2000...	0
5170	Subject: important online banking alert\r\ndea...	1

รูปที่ 3.5 ตัวอย่างชุดข้อมูลจดหมายอิเล็กทรอนิกส์ก่อนเปลี่ยนชื่อสดมภ์ label_num เป็น label

จากรูปที่ 3.5 ทำการเปลี่ยนชื่อตัวแปร label_num เพื่อให้เข้าใจง่ายขึ้น

	text	label
0	Subject: enron methanol ; meter # : 988291\r\n...	0
1	Subject: hpl nom for january 9 , 2001\r\n(see...	0
2	Subject: neon retreat\r\nho ho ho , we ' re ar...	0
3	Subject: photoshop , windows , office . cheap ...	1
4	Subject: re : indian springs\r\nthis deal is t...	0
...	...	...
5166	Subject: put the 10 on the ft\r\nthe transport...	0
5167	Subject: 3 / 4 / 2000 and following noms\r\nhp...	0
5168	Subject: calpine daily gas nomination\r\n>\r\n...	0
5169	Subject: industrial worksheets for august 2000...	0
5170	Subject: important online banking alert\r\ndea...	1

รูปที่ 3.6 ตัวอย่างชุดข้อมูลจดหมายอิเล็กทรอนิกส์หลังเปลี่ยนชื่อสดมภ์ label_num เป็น label

จากรูปที่ 3.6 เปลี่ยนชื่อตัวแปรเป็น label เพื่อให้เข้าใจง่ายในการวิเคราะห์ข้อมูล

3.3.2.3 การสร้างสดมภ์ใหม่

ใช้คำสั่ง DataFrame จากชุดคำสั่ง pandas ในการสร้างสดมภ์ใหม่ชื่อ CleanedText เพื่อเก็บข้อความของจดหมายอิเล็กทรอนิกส์ที่จะถูกนำไปทำความสะอาดต่อไป ดังรูปที่ 3.7

	text	label	CleanedText
0	Subject: enron methanol ; meter # : 988291\r\n...	0	Subject: enron methanol ; meter # : 988291\r\n...
1	Subject: hpl nom for january 9 , 2001\r\n(see...	0	Subject: hpl nom for january 9 , 2001\r\n(see...
2	Subject: neon retreat\r\nho ho ho , we ' re ar...	0	Subject: neon retreat\r\nho ho ho , we ' re ar...
3	Subject: photoshop , windows , office . cheap ...	1	Subject: photoshop , windows , office . cheap ...
4	Subject: re : indian springs\r\nthis deal is t...	0	Subject: re : indian springs\r\nthis deal is t...
...	...	...	...
5166	Subject: put the 10 on the ft\r\nthe transport...	0	Subject: put the 10 on the ft\r\nthe transport...
5167	Subject: 3 / 4 / 2000 and following noms\r\nhp...	0	Subject: 3 / 4 / 2000 and following noms\r\nhp...
5168	Subject: calpine daily gas nomination\r\n>\r\n...	0	Subject: calpine daily gas nomination\r\n>\r\n...
5169	Subject: industrial worksheets for august 2000...	0	Subject: industrial worksheets for august 2000...
5170	Subject: important online banking alert\r\ndea...	1	Subject: important online banking alert\r\ndea...

รูปที่ 3.7 ตัวอย่างชุดข้อมูลจดหมายอิเล็กทรอนิกส์หลังสร้างสดมภ์ใหม่ชื่อ CleanedText

จากรูปที่ 3.7 ทำการสร้างสดมภ์ใหม่ชื่อ CleanedText เพื่อไว้เก็บข้อมูลที่ผ่านการจัดเตรียมข้อมูลมาแล้ว

3.3.3 การลบหัวเรื่องของจดหมายอิเล็กทรอนิกส์

ใช้คำสั่ง replace ในการลบหัวเรื่องของจดหมายอิเล็กทรอนิกส์ชื่อ Subject ออกจากข้อความไม่มีผลต่อผลวิเคราะห์ เนื่องจากเป็นคำที่อยู่หน้าข้อความในรูปแบบเดียวกันทุกข้อความ ซึ่งการลบ Subject หน้าข้อความออกจะทำให้ลดเวลาในการประมวลผลลงได้ ตัวอย่างชุดข้อมูลจดหมายอิเล็กทรอนิกส์หลังจากลบ Subject ออกจากข้อความ แสดงดังรูปที่ 3.8 และ 3.9

	text	label	CleanedText
0	Subject: enron methanol ; meter # : 988291\r\n...	0	Subject: enron methanol ; meter # : 988291\r\n...
1	Subject: hpl nom for january 9 , 2001\r\n(see...	0	Subject: hpl nom for january 9 , 2001\r\n(see...
2	Subject: neon retreat\r\nho ho ho , we ' re ar...	0	Subject: neon retreat\r\nho ho ho , we ' re ar...
3	Subject: photoshop , windows , office . cheap ...	1	Subject: photoshop , windows , office . cheap ...
4	Subject: re : indian springs\r\nthis deal is t...	0	Subject: re : indian springs\r\nthis deal is t...
...	...	...	...
5166	Subject: put the 10 on the ft\r\nthe transport...	0	Subject: put the 10 on the ft\r\nthe transport...
5167	Subject: 3 / 4 / 2000 and following noms\r\nhp...	0	Subject: 3 / 4 / 2000 and following noms\r\nhp...
5168	Subject: calpine daily gas nomination\r\n>\r\n...	0	Subject: calpine daily gas nomination\r\n>\r\n...
5169	Subject: industrial worksheets for august 2000...	0	Subject: industrial worksheets for august 2000...
5170	Subject: important online banking alert\r\ndea...	1	Subject: important online banking alert\r\ndea...

รูปที่ 3.8 ตัวอย่างชุดข้อมูลจดหมายอิเล็กทรอนิกส์ก่อนลบ Subject ออกจากข้อความ

จากรูปที่ 3.8 เนื่องจากข้อมูลที่นำมามีข้อความ Subject ซึ่งหมายถึงเรื่อง เป็นคำที่ไม่ได้นำมาใช้วิเคราะห์จึงต้องทำการลบข้อความ Subject ออก

	text	label	CleanedText
0	Subject: enron methanol ; meter # : 988291\r\n...	0	: enron methanol ; meter # : 988291\r\nthis is...
1	Subject: hpl nom for january 9 , 2001\r\n(see...	0	: hpl nom for january 9 , 2001\r\n(see attach...
2	Subject: neon retreat\r\nho ho ho , we ' re ar...	0	: neon retreat\r\nho ho ho , we ' re around to...
3	Subject: photoshop , windows , office . cheap ...	1	: photoshop , windows , office . cheap . main ...
4	Subject: re : indian springs\r\nthis deal is t...	0	: re : indian springs\r\nthis deal is to book ...
...	...	...	...
5166	Subject: put the 10 on the ft\r\nthe transport...	0	: put the 10 on the ft\r\nthe transport volume...
5167	Subject: 3 / 4 / 2000 and following noms\r\nhp...	0	: 3 / 4 / 2000 and following noms\r\nhpl can '...
5168	Subject: calpine daily gas nomination\r\n>\r\n...	0	: calpine daily gas nomination\r\n>\r\n>\r\nju...
5169	Subject: industrial worksheets for august 2000...	0	: industrial worksheets for august 2000 activi...
5170	Subject: important online banking alert\r\ndea...	1	: important online banking alert\r\ndear value...

รูปที่ 3.9 ตัวอย่างชุดข้อมูลจดหมายอิเล็กทรอนิกส์หลังลบ Subject ออกจากข้อความ

จากรูปที่ 3.9 หลังจากลบคำว่า Subject ออก ข้อความที่ลบจะไปถูกจัดเก็บไว้ที่สดมภ์ CleanedText จะเห็นได้ว่าในกรอบสี่เหลี่ยมสีแดง คำว่า Subject จะถูกลบไปเพื่อที่จะนำไปใช้วิเคราะห์

3.3.4 การเปลี่ยนตัวอักษรพิมพ์ใหญ่เป็นพิมพ์เล็ก

ใช้คำสั่ง str.lower() ในการเปลี่ยนตัวอักษรพิมพ์ใหญ่ให้เป็นตัวอักษรพิมพ์เล็กทั้งหมด เนื่องจากตัวอักษรพิมพ์เล็กและพิมพ์ใหญ่มีผลต่อการตรวจจับของคอมพิวเตอร์

3.3.5 การลบคำฟุ่มเฟือย (Stop Word)

ใช้ชุดคำสั่ง stopwords ในการลบคำที่เป็นคำฟุ่มเฟือยออก เนื่องจากเป็นคำที่ไม่สื่อความหมาย โดยใช้คลังคำศัพท์ภาษาอังกฤษชื่อ stopword ในชุดคำสั่งของ nltk เพื่อให้การตัดคำภาษาอังกฤษมีความถูกต้องแม่นยำมากที่สุด ตัวอย่างคำฟุ่มเฟือยที่ถูกลบ แสดงดังรูปที่ 3.10

```
"they're",
"they've",
'this',
'those',
'through',
'to',
'too',
'under',
'until',
'up',
'very',
'was',
"wasn't",
'we',
"we'd",
"we'll",
"we're",
```

รูปที่ 3.10 ตัวอย่างคำฟุ่มเฟือย

จากรูปที่ 3.10 เป็นตัวอย่างคำฟุ่มเฟือยที่พบในจดหมายอิเล็กทรอนิกส์ จึงทำการลบคำฟุ่มเฟือยออกจากข้อความเพื่อลดคำที่ไม่มีความหมายออก แล้วนำไปใช้วิเคราะห์

3.3.6 การลบลิงก์ที่นำไปสู่เว็บไซต์

ใช้คำสั่ง sub จากชุดคำสั่ง re ในการลบลิงก์ที่นำไปสู่เว็บไซต์ เนื่องจากชื่อลิงก์ที่นำไปสู่เว็บไซต์อาจเป็นคำที่ไม่มีความหมาย หากไม่ลบลิงก์ออกก่อนจะทำการตัดคำภาษาอังกฤษเกิดความผิดพลาดและส่งผลต่อผลการวิเคราะห์ได้ ตัวอย่างชุดข้อมูลจดหมายอิเล็กทรอนิกส์หลังจากลบลิงก์ที่นำไปสู่เว็บไซต์ แสดงดังรูปที่ 3.11 และ 3.12

Subject: looking medication ? ` best source . difficult make material condition better best law , easy enough ruin bad laws . excuse . . . : ) found best simpliest site medication net . perscription , easy delivery . private , secure , easy . better see rightly pound week squint million . ` got anything ever want . erection treatment pills . anti - depressant pills , weight loss , ! http : / / splicings . bombahakcx . com / 3 / knowledge human power synonymous . high - quality stuff low rates ! 100 % moneyback guarantee ! god , nature sufficeth unto wise hath need author .

รูปที่ 3.11 ตัวอย่างชุดข้อมูลจดหมายอิเล็กทรอนิกส์ก่อนลบลิงก์ที่นำไปสู่เว็บไซต์

จากรูปที่ 3.11 เนื่องจากในข้อความของจดหมายอิเล็กทรอนิกส์มีลิงก์ที่นำไปสู่เว็บไซต์ ซึ่งอาจทำให้การวิเคราะห์ข้อมูลผิดพลาดได้จึงต้องทำการลบลิงก์ที่นำไปสู่เว็บไซต์ออก

looking medication?` best source.difficult make material condition better best law , easy enough ruin bad laws.excuse... ) found best simpliest site medication net.perscription , easy delivery.private , secure , easy.better see rightly pound week squint million.` got anything ever want.erection treatment pills , anti-depressant pills , weight loss , ! knowledge human power synonymous.high-quality stuff low rates ! % moneyback guarantee ! god , nature sufficeth unto wise hath need author .

รูปที่ 3.12 ตัวอย่างชุดข้อมูลจดหมายอิเล็กทรอนิกส์หลังลบลิงก์ที่นำไปสู่เว็บไซต์

จากรูปที่ 3.12 ทำการลบลิงก์ที่นำไปสู่เว็บไซต์ในข้อความจดหมายอิเล็กทรอนิกส์เพื่อช่วยให้การวิเคราะห์มีความถูกต้องมากยิ่งขึ้น

3.3.7 การลบตัวอักขระพิเศษ (Special Character)

ใช้คำสั่ง sub และคำสั่ง replace ในการลบตัวอักขระพิเศษ เนื่องจากชุดข้อมูลจดหมายอิเล็กทรอนิกส์ในสดมภ์ text เป็นข้อความที่มีตัวอักขระพิเศษอยู่ เช่น ! - # / 0-9 จึงต้องทำการลบตัวอักขระพิเศษออกจากข้อความ หากไม่ทำการตัดตัวอักขระพิเศษอาจทำให้การวิเคราะห์ข้อมูลเกิดความผิดพลาดได้ ตัวอย่างชุดข้อมูลจดหมายอิเล็กทรอนิกส์หลังจากลบตัวอักขระพิเศษ แสดงดังรูปที่ 3.13

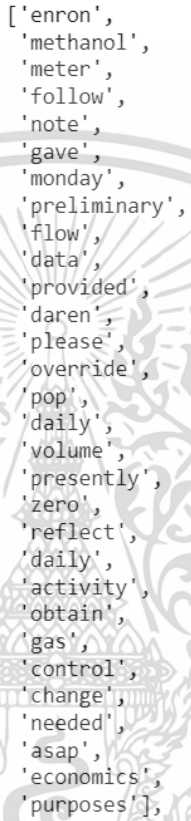
	text	label	CleanedText
0	Subject: enron methanol ; meter # : 988291\r\n...	0	enron methanol meter follow note gave monday p...
1	Subject: hpl nom for january 9 , 2001\r\n(see...	0	hpl nom january see attached file hplnol xls h...
2	Subject: neon retreat\r\nho ho ho , we ' re ar...	0	neon retreat ho ho ho around wonderful time ye...
3	Subject: photoshop , windows , office , cheap ...	1	photoshop windows office cheap main trending a...
4	Subject: re : indian springs\r\nthis deal is t...	0	indian springs deal book teco pvr revenue unde...
...	...	...	...
5166	Subject: put the 10 on the ft\r\nthe transport...	0	put ft transport volumes decreased contract th...
5167	Subject: 3 / 4 / 2000 and following noms\r\nhp...	0	following noms hpl take extra mmcf weekend try...
5168	Subject: calpine daily gas nomination\r\n>\r\n...	0	calpine daily gas nomination julie mention ear...
5169	Subject: industrial worksheets for august 2000...	0	industrial worksheets august activity attached...
5170	Subject: important online banking alert\r\ndea...	1	important online banking alert dear valued cit...

รูปที่ 3.13 ตัวอย่างชุดข้อมูลจดหมายอิเล็กทรอนิกส์ที่ลบตัวอักขระพิเศษออกจากข้อความ

จากรูปที่ 3.13 เนื่องจากตัวเลขไม่ได้เป็นข้อความ จึงทำการลบตัวอักขระพิเศษในข้อความจดหมายอิเล็กทรอนิกส์ เพื่อใช้ในการวิเคราะห์ข้อมูล หลังจากที่ลบตัวอักขระพิเศษแล้ว ข้อความที่ลบจะไปถูกจัดเก็บไว้ในสดมภ์ CleanedText

3.3.8 การตัดข้อความออกเป็นคำ (Word Tokenization)

ใช้คำสั่ง split() ในการนำข้อความที่ถูกทำความสะอาดแล้วในสดมภ์ CleanedText มาตัดแบ่งออกเป็นคำๆ เนื่องจากข้อความในงานวิจัยนี้เป็นภาษาอังกฤษ จึงสามารถใช้หลักการตัดคำตามช่องว่างได้เลย ตัวอย่างคำศัพท์หลังตัดข้อความจดหมายอิเล็กทรอนิกส์ออกเป็นคำ แสดงดังรูปที่ 3.14



```
['enron',
 'methanol',
 'meter',
 'follow',
 'note',
 'gave',
 'monday',
 'preliminary',
 'flow',
 'data',
 'provided',
 'daren',
 'please',
 'override',
 'pop',
 'daily',
 'volume',
 'presently',
 'zero',
 'reflect',
 'daily',
 'activity',
 'obtain',
 'gas',
 'control',
 'change',
 'needed',
 'asap',
 'economics',
 'purposes'],
```

รูปที่ 3.14 ตัวอย่างคำศัพท์หลังตัดข้อความจดหมายอิเล็กทรอนิกส์ออกเป็นคำ

จากรูปที่ 3.14 ทำการตัดข้อความในจดหมายอิเล็กทรอนิกส์ออกเป็นคำ เพื่อให้สามารถนำใช้ใช้ในการสร้างเวกเตอร์ของคำด้วย Word2Vec ได้

3.3.9 การสร้างเวกเตอร์ของคำด้วย Word2Vec

ขั้นตอนนี้จะเป็นการสร้างแบบจำลองที่ชื่อว่า Word2Vec สร้างการฝังคำหรือแปลงคำให้อยู่ในรูปแบบของเวกเตอร์ ซึ่งเวกเตอร์ของคำต่างๆ ถูกคำนวณจากบริบทรอบข้าง มีหลักการในการเปรียบเทียบเวกเตอร์ทางความหมายของคำทั้ง 2 กลุ่ม แล้วคืนค่าออกมาเป็นตัวเลขตั้งแต่ -1 ถึง 1 บ่งชี้ถึงความใกล้เคียงทางความหมายโดยให้ค่าจากน้อยไปมาก หรือพูดอีกนัยหนึ่งว่าคำที่มีบริบทการปรากฏคล้ายๆ กันควรเป็นคำที่มีความหมายคล้ายกันด้วย ซึ่งวิธี Word2vec สามารถวัดค่าความคล้ายทางความหมายของเวกเตอร์ของคำและใช้ร่วมกับการจำแนกกลุ่มข้อมูลข้อความได้ดี (บุษบงก์และคณะ, 2564) โดยคณะผู้วิจัยสร้างแบบจำลอง Word2vec จากชุดคำสั่ง Gensim และนำค่า

เวกเตอร์ที่ได้ไปใช้สร้างแบบจำลองในการกำจัดค่านอกเกณฑ์ต่อไป ตัวอย่างชุดข้อมูลจดหมายอิเล็กทรอนิกส์หลังแปลงคำให้อยู่ในรูปแบบของเวกเตอร์ แสดงดังรูปที่ 3.15

	Word	0	1	2	3	4	5	6	7	8	...
0	enron	3.928696	0.504737	0.013813	1.373888	0.258697	1.588908	1.864040	-0.511628	-0.591747	...
1	methanol	1.120708	-0.317901	-0.278764	0.886773	-0.154468	0.263497	0.488889	-0.637664	0.310296	...
2	meter	2.584731	-0.344332	-0.111579	2.107401	0.191587	0.205657	0.791825	-2.095314	0.760698	...
3	follow	0.932220	0.005785	0.077527	0.389138	-0.256781	0.461640	-0.075094	0.770568	0.163480	...
4	note	1.394642	-0.214695	0.447582	0.936922	-0.307532	0.523434	-0.396522	0.280574	0.441946	...
...	...	...	...	...	...	...	...	...	...	...	...
45082	macintoshdogleg	0.022022	-0.014079	0.001009	0.012649	-0.004412	0.018012	0.016358	0.011056	0.011932	...
45083	ilaa	0.003804	-0.007655	-0.000902	0.003940	-0.003541	0.001738	0.003795	0.004320	0.004855	...
45084	liqaa	0.006072	-0.000719	0.003521	-0.001763	-0.004332	0.004709	0.001434	0.007712	0.003923	...
45085	buydeaug	0.007383	0.003751	0.037424	0.015994	0.023972	-0.006834	-0.017004	-0.013930	0.005442	...
45086	charterr	0.000290	0.012084	-0.004488	-0.001725	-0.010732	0.006879	0.000386	0.004669	-0.005080	...

รูปที่ 3.15 ตัวอย่างชุดข้อมูลจดหมายอิเล็กทรอนิกส์ที่แปลงคำให้อยู่ในรูปแบบของเวกเตอร์

จากรูปที่ 3.15 ทำการสร้างเวกเตอร์ของคำด้วย Word2Vec โดยใช้คำในจดหมายอิเล็กทรอนิกส์ซึ่งค่าเวกเตอร์ที่ได้มีขนาด 100 เวกเตอร์ต่อคำ 1 คำ

3.3.10 การลดมิติเวกเตอร์ของคำโดยใช้การวิเคราะห์ส่วนประกอบหลัก (Principal Component Analysis)

หลักจากทำการสร้างเวกเตอร์ของคำทุกๆ คำโดยการใช้แบบจำลอง Word2vec แล้วจะเห็นได้ว่าคำแต่ละคำประกอบไปด้วยเวกเตอร์เป็นจำนวนมาก ทำให้การแสดงผลของข้อมูลทำได้ยาก คณะผู้วิจัยจึงทำการลดมิติของเวกเตอร์แต่ละคำโดยใช้การวิเคราะห์ส่วนประกอบหลัก (Principal Component Analysis) ซึ่งเป็นวิธีการหาความสัมพันธ์ระหว่างตัวแปรของข้อมูลว่ามีรูปแบบเป็นยังไง เพื่อนำไปทำกระบวนการลดมิติของข้อมูลโดยที่ไม่ทำให้สูญเสียข้อมูลสำคัญไป (ภัครพล, 2564) คณะผู้วิจัยทำการลดมิติเวกเตอร์ของคำโดยใช้คำสั่ง PCA จากชุดคำสั่ง sklearn แล้วกำหนดจำนวนมิติเวกเตอร์ของคำเท่ากับ 2 เวกเตอร์ ให้เป็นตัวแปร x และ y ทำให้สามารถแสดงผลเป็นกราฟเพื่อง่ายต่อการวิเคราะห์ข้อมูลและสามารถสร้างแบบจำลองกำจัดค่านอกเกณฑ์ต่อไป ตัวอย่างชุดข้อมูลจดหมายอิเล็กทรอนิกส์หลังทำการลดมิติโดยใช้การวิเคราะห์ส่วนประกอบหลักแสดงดังรูปที่ 3.16

	Word	x	y
0	enron	7.868763	10.162171
1	methanol	3.068713	1.349805
2	meter	8.838520	6.330078
3	follow	3.924496	-1.414053
4	note	4.699094	-0.578436
...	...	...	...
45082	macintoshdogleg	-0.394645	-0.012005
45083	ilaa	-0.431038	0.013621
45084	liqaa	-0.455730	-0.002274
45085	buydeaug	-0.576120	0.049216
45086	charterr	-0.509459	-0.030523

รูปที่ 3.16 ตัวอย่างชุดข้อมูลจดหมายอิเล็กทรอนิกส์หลังทำการลดมิติโดยใช้การวิเคราะห์ส่วนประกอบหลัก

จากรูปที่ 3.16 เนื่องจากการสร้างเวกเตอร์ของคำด้วย Word2Vec ได้ 100 เวกเตอร์ซึ่งมันมีขนาดใหญ่เลยทำให้ใช้เวลาในการประมวลผลที่นานจึงได้ทำการลดมิติเวกเตอร์ของคำให้เหลือเพียง 2 เวกเตอร์

3.4 การกำจัดค่านอกเกณฑ์ (Outlier Detection)

หลังจากทำการลดมิติเวกเตอร์ของคำโดยใช้การวิเคราะห์ส่วนประกอบหลักแล้ว จึงได้นำเวกเตอร์ที่ได้ไปใช้หาค่านอกเกณฑ์ทั้งหมด 3 วิธี ได้แก่ วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว (Isolation Forest) และวิธี IF-LOF โดยจะนำคำที่เป็นค่านอกเกณฑ์ที่ได้ในแต่ละวิธีมากำจัดออกจากข้อความที่ผ่านขั้นตอนการจัดเตรียมข้อมูล

3.4.1 วิธี DBSCAN

วิธี DBSCAN ทำการหาค่านอกเกณฑ์โดยหาบริเวณที่ข้อมูลเกาะกลุ่มกันเป็นจำนวนมากซึ่งสามารถคำนวณได้จากรัศมีบริเวณรอบ ๆ ของจุดข้อมูลที่กำหนด ซึ่งการที่จะใช้วิธี DBSCAN ได้จำเป็นต้องมีพารามิเตอร์ 2 ตัว คือ รัศมีจากจุดศูนย์กลาง (Epsilon) และจำนวนจุดข้อมูลขั้นต่ำ (Min Points) สำหรับการกำหนดจุดศูนย์กลาง (Center) คณะผู้วิจัยทำการสร้างแบบจำลอง DBSCAN จากชุดคำสั่ง sklearn

3.4.2 วิธีป่าไม้โดดเดี่ยว (Isolation Forest)

วิธีป่าไม้โดดเดี่ยวเพื่อหาค่านอกเกณฑ์ คำนวณเป็นคะแนนความผิดปกติ (Anomaly score) เพื่อใช้ในการแยกประเภทของข้อมูลได้ ซึ่งคะแนนความความผิดปกตินั้นจะมีค่าตั้งแต่ 0 ถึง 1 โดยยิ่งค่าน้อยจะหมายถึงข้อมูลที่มีแนวโน้มผิดปกติ ส่วนข้อมูลที่มีค่ามากกว่า 0.5 ขึ้นไปจะถือว่าเป็นข้อมูลทั่วไปที่ไม่มีความผิดปกติ โดยคณะผู้วิจัยสร้างแบบจำลอง IsolationForest จากชุดคำสั่ง sklearn ในการสร้างแบบจำลองป่าไม้โดดเดี่ยว

3.4.3 วิธี IF-LOF (Isolation Forest- Local Outlier Factor)

วิธี IF-LOF เป็นวิธีการเรียนรู้แบบรวมกลุ่ม เนื่องจากวิธีป่าไม้โดดเดี่ยวมีความอ่อนแอต่อการจัดการกับค่านอกเกณฑ์ในพื้นที่ (Local Outlier) ในขณะที่วิธี LOF ทำงานได้ดีในการตรวจจับค่านอกเกณฑ์ในพื้นที่ แต่ใช้เวลานานในการประมวลผล (Zhangyu et al, 2019) โดยขั้นตอนการทำงานของวิธี IF-LOF จะทำการนำค่านอกเกณฑ์ที่ได้จากการสร้างแบบจำลองวิธีป่าไม้โดดเดี่ยว มาใช้ในการสร้างแบบจำลอง LOF โดยคณะผู้วิจัยทำการสร้างแบบจำลอง LOF จากแบบจำลอง LocalOutlierFactor ในชุดคำสั่ง sklearn

การกำจัดคำที่เป็นค่านอกเกณฑ์ออกจากข้อความขั้นตอนนี้จะเป็นการนำคำที่เป็นค่านอกเกณฑ์ของแต่ละวิธีมากำจัดออกจากข้อความที่ผ่านขั้นตอนการจัดเตรียมข้อมูล จากนั้นสร้างสดมภ์เพื่อเก็บข้อความที่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์ของวิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF เพื่อนำไปใช้เป็นข้อมูลเข้าในการสร้างแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกต่อไป

3.5 การแบ่งข้อมูล

ขั้นตอนนี้คณะผู้วิจัยได้นำข้อความทั้งที่กำจัดค่านอกเกณฑ์ และไม่ได้กำจัดค่านอกเกณฑ์มาแบ่งข้อมูลออกเป็นชุดข้อมูลฝึกสอนและชุดข้อมูลทดสอบในอัตราส่วน 70:30 กล่าวคือ จำนวนข้อมูลที่ใช้ในการสร้างแบบจำลองมี 3,619 ตัว และจำนวนข้อมูลที่ใช้ทดสอบ 1,552 ตัว แล้วนำชุดข้อมูลฝึกสอนไปสร้างแบบจำลองการเรียนรู้ของเครื่อง และใช้ชุดข้อมูลทดสอบในการประเมินประสิทธิภาพของเบบจำลอง

3.6 การเรียนรู้ของเครื่อง (Machine Learning)

ขั้นตอนต่อมาคือการสร้างแบบจำลองการเรียนรู้ของเครื่อง (Machine Learning) เป็นการทำให้ระบบคอมพิวเตอร์เรียนรู้และสร้างขั้นตอนวิธี (Algorithm) ที่สามารถเรียนรู้ข้อมูลและทำนายข้อมูลได้ (วีระพันธ์, 2564) ซึ่งคณะผู้วิจัยได้เลือกใช้การเรียนรู้ของเครื่องทั้งหมด 3 วิธี ได้แก่ วิธีนาอีฟเบส์ วิธีเพื่อนบ้านใกล้สุด k ตัว และวิธีซัพพอร์ตเวกเตอร์แมชชีน โดยนำสดมภ์ที่เก็บข้อความที่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์ของวิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF และไม่ได้กำจัดค่านอกเกณฑ์ ในชุดข้อมูลฝึกสอน ซึ่งประกอบได้ด้วยข้อความของจดหมายอิเล็กทรอนิกส์และหมายเลขคำตอบของข้อมูลไปใช้เป็นข้อมูลเข้าในการสร้างแบบจำลองการเรียนรู้ของเครื่อง

3.6.1 วิธีนาอีฟเบส์ (Naive Bayes)

โดยนำข้อมูลที่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์ของวิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF มาใช้เป็นข้อมูลเข้าซึ่งคณะผู้วิจัยได้สร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีนาอีฟเบส์จากแบบจำลอง MultinomialNB ในชุดคำสั่ง sklearn ในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์

3.6.2 วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machaine)

โดยนำข้อมูลที่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์ของวิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF มาใช้เป็นข้อมูลเข้าในการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์ คณะผู้วิจัยใช้แบบจำลอง svm ในชุดคำสั่ง sklearn ในการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ฟังก์ชันเคอร์เนลที่นำมาใช้คือฟังก์ชันเกาส์เซียนเรเดียลเบสิส (Gaussian Radial Basis Function-RBF) คือ การเรียนรู้ปรับค่าน้ำหนักให้ได้ฟังก์ชันการส่งที่เหมาะสมที่สุด (กาญจนา, 2561)

3.6.3 วิธีเพื่อนบ้านใกล้สุด k ตัว (K-Nearest Neighbors)

โดยนำข้อมูลที่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์ของวิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF มาใช้เป็นข้อมูลเข้าในการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้วิธีระยะห่างยุคลิด (Euclidian Distance) ในการกำหนดค่า k ไม่มีกฎตายตัวขึ้นอยู่กับข้อมูล โดยจะทำการพิจารณาค่า k ที่ปรับเปลี่ยนในแต่ละค่าเปรียบเทียบกับประสิทธิภาพค่าความแม่นของผลการทดสอบ (สุจิรา, 2560) ผู้วิจัยทำการพิจารณาค่า k ตั้งแต่ 1 ถึง 15 และใช้แบบจำลอง KneighborsClassifier จากชุดคำสั่ง sklearn ในการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์

3.7 การเรียนรู้เชิงลึก (Deep Learning)

ขั้นตอนต่อมาคือการสร้างแบบจำลองการเรียนรู้เชิงลึก (Deep Learning) เป็นปัญญาประดิษฐ์ (AI) รูปแบบหนึ่ง ที่เลียบแบบกระบวนการประมวลผลในสมองมนุษย์ที่มีความซับซ้อน (ณัฏฐณิชา, 2562) ซึ่งคณะผู้วิจัยได้เลือกใช้การเรียนรู้เชิงลึกทั้งหมด 3 วิธี ได้แก่วิธีโครงข่ายประสาทเทียม วิธีโครงข่ายแบบวนซ้ำ และวิธีหน่วยความจำระยะสั้นยาวโดยนำสดมภ์ที่เก็บข้อความที่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์ของวิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF และไม่ได้กำจัดค่านอกเกณฑ์ในชุดข้อมูลฝึกสอน ซึ่งประกอบได้ด้วยข้อความของจดหมายอิเล็กทรอนิกส์และหมายเลขคำตอบของข้อมูลไปใช้เป็นข้อมูลเข้าในการสร้างแบบจำลองการเรียนรู้เชิงลึก

3.7.1 วิธีโครงข่ายประสาทเทียม (Artificial Neural Network)

โดยนำข้อมูลที่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์ของวิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF มาใช้เป็นข้อมูลเข้า ในการสร้างแบบจำลองวิธีโครงข่ายประสาทเทียมจะประกอบไปด้วย ชั้นข้อมูลเข้า ชั้นซ่อนและชั้นข้อมูลออก ซึ่งจะช่วยทำให้คอมพิวเตอร์ทำงานเหมือนมนุษย์ มีการจดจำรูปแบบและสร้างองค์ความรู้ใหม่ ในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์

ซึ่งโครงสร้างประกอบด้วย ชั้นข้อมูลเข้า ชั้นซ่อน และชั้นข้อมูลออก ภายในแต่ละชั้น จะประกอบด้วยโหนด ซึ่งความซับซ้อนของจำนวนชั้นและโหนดขึ้นอยู่กับการออกแบบและความเหมาะสมในการทำงานรวมทั้งการทดสอบผล (วิศรุต และวริศ, 2563) โดยใช้ชุดคำสั่ง TensorFlow และ Keras ในการสร้างแบบจำลอง

ตารางที่ 3.2 ค่าพารามิเตอร์ที่ใช้ในการการเรียนรู้เชิงลึกด้วยวิธีโครงข่ายประสาทเทียม

พารามิเตอร์	
Hidden layer	1
Activation function	relu, sigmoid
Optimizer	Adam
loss	binary_crossentropy
input_length	3338,2588,2420,3235

จากตารางที่ 3.2 คณะผู้วิจัยได้ใช้ชั้นซ่อนที่ 1 ชั้นเริ่มจากนำข้อมูลเข้าของข้อความทั้งที่กำจัดค่านอกเกณฑ์และไม่ได้กำจัดค่านอกเกณฑ์ โดยที่ข้อมูลเข้าของข้อความที่กำจัดค่านอกเกณฑ์ทั้ง 3 วิธี ได้แก่ วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF เท่ากับ 2,588, 2,420, 3,235 ตามลำดับ และ ข้อมูลเข้าของข้อความทั้งที่ไม่กำจัดค่านอกเกณฑ์เท่ากับ 3,338 ต่อมาชั้นซ่อนใช้เพื่อเรียนรู้ข้อมูลซึ่งใช้ฟังก์ชั่นกระตุ้นเป็น relu และในชั้นข้อมูลออกได้ใช้ฟังก์ชั่นกระตุ้นเป็น sigmoid เพื่อทำให้ข้อมูล

ออกมี จำนวน Node เท่ากับ 2 โดยใช้อัตราการเรียนรู้เป็น Adam ซึ่งเป็นการปรับค่าอัตราการเรียนรู้ให้ เหมาะสมจากนั้นได้ใช้การหาค่าสูญหายเป็น binary_crossentropy

3.7.2 วิธีโครงข่ายประสาทแบบวนซ้ำ (Recurrent Neural Network)

โดยนำข้อมูลที่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์ของวิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF มาใช้เป็นข้อมูลเข้า โดยใช้ชุดคำสั่ง TensorFlow และ Keras ในการสร้างแบบจำลอง ซึ่งเป็นการเอาผลลัพธ์ที่ได้จากชั้นซ่อนก่อนหน้ามาใช้เป็นข้อมูลเข้าในชั้นซ่อนปัจจุบัน ซึ่งนำมาเป็นข้อมูลเข้าใหม่คู่กับข้อมูลเข้าแบบปกติ ซึ่งจะช่วยให้แบบจำลองมีการจำรูปแบบของลำดับข้อมูลเข้าในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์

ตารางที่ 3.3 ค่าพารามิเตอร์ที่ในการการเรียนรู้เชิงลึกด้วยวิธีโครงข่ายประสาทแบบวนซ้ำ

พารามิเตอร์	
Hidden layer	1
Activation function	tanh, sigmoid
Optimizer	Adam
loss	binary_crossentropy
input_length	3,338, 2,588, 2,420, 3,235

จากตารางที่ 3.3 คณะผู้วิจัยได้ใช้ชั้นซ่อนที่ 1 ชั้นเริ่มจากนำข้อมูลเข้าของข้อความทั้งที่กำจัดค่านอกเกณฑ์และไม่ได้กำจัดค่านอกเกณฑ์ โดยที่ข้อมูลเข้าของข้อความที่กำจัดค่านอกเกณฑ์ทั้ง 3 วิธี ได้แก่ วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF เท่ากับ 2,588, 2,420, 3,235 ตามลำดับ และ ข้อมูลเข้าของข้อความทั้งที่ไม่กำจัดค่านอกเกณฑ์เท่ากับ 3,338 ต่อมาชั้นซ่อนใช้เพื่อเรียนรู้ข้อมูล ซึ่งใช้ฟังก์ชั่นกระตุ้นเป็น tanh และในชั้นข้อมูลออกได้ใช้ฟังก์ชั่นกระตุ้นเป็น sigmoid เพื่อทำให้ข้อมูลออกมี จำนวน Node เท่ากับ 2 โดยใช้อัตราการเรียนรู้เป็น Adam ซึ่งเป็นการปรับค่าอัตราการเรียนรู้ให้ เหมาะสมจากนั้นได้ใช้การหาค่าสูญหายเป็น binary_crossentropy

3.7.3 วิธีหน่วยความจำระยะสั้นยาว (Long Short-Term Memory)

วิธีหน่วยความจำระยะสั้นยาวเป็นโครงข่ายประสาทแบบวนซ้ำ รูปแบบหนึ่งซึ่งได้มีการพัฒนาให้มีความเสถียรมากขึ้น โดยวิธีนี้สามารถเก็บสถานะ หรือข้อมูลในแต่ละโหนดเพื่อที่เวลาย้อนไป ดูจะได้ทราบว่าข้อมูลค่าเดิมมาจากค่าอะไร และจุดเด่นของแบบจำลองนี้คือ ฟังก์ชันที่เปรียบเสมือนประตูซึ่งประกอบด้วยประตูลืม ประตูข้อมูลเข้า และประตูข้อมูลออก โดยจะคอยคุมข้อมูลที่จะเข้ามาในแต่ละโหนด (วิศรุต และวริศ, 2563) ใช้ชุดคำสั่ง TensorFlow และ Keras ในการสร้างแบบจำลอง

ตารางที่ 3.4 ค่าพารามิเตอร์ที่ในการการเรียนรู้เชิงลึกด้วยวิธีหน่วยความจำระยะสั้นยาว

พารามิเตอร์	
Hidden layer	1
Activation function	relu, sigmoid
Optimizer	Adam
loss	binary_crossentropy
input_length	3,338, 2,588, 2,420, 3,235

จากตารางที่ 3.4 คณะผู้วิจัยได้ใช้ชั้นซ่อนที่ 1 ชั้นเริ่มจากนำข้อมูลเข้าของข้อความทั้งที่กำจัดค่านอกเกณฑ์และไม่ได้กำจัดค่านอกเกณฑ์ โดยที่ข้อมูลเข้าของข้อความที่กำจัดค่านอกเกณฑ์ทั้ง 3 วิธี ได้แก่ วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF เท่ากับ 2,588, 2,420, 3,235 ตามลำดับ และ ข้อมูลเข้าของข้อความทั้งที่ไม่กำจัดค่านอกเกณฑ์เท่ากับ 3,338 ต่อมาชั้นซ่อนใช้เพื่อเรียนรู้ข้อมูล ซึ่งใช้ฟังก์ชั่นกระตุ้นเป็น relu และในชั้นข้อมูลออกได้ใช้ฟังก์ชั่นกระตุ้นเป็น sigmoid เพื่อทำให้ข้อมูลออกมี จำนวน Node เท่ากับ 2 โดยใช้อัตราการเรียนรู้เป็น Adam ซึ่งเป็นการปรับค่าอัตราการเรียนรู้ให้ เหมาะสมจากนั้นได้ใช้การหาค่าสูญหายเป็น binary_crossentropy

3.8 การเรียนรู้แบบรวมกลุ่ม (Ensemble Learning)

แบบจำลองการเรียนรู้แบบรวมกลุ่มเป็นการสร้างแบบจำลองที่มีการใช้ตัวจำแนก (Classifier) มากกว่าหนึ่งตัวในการเรียนรู้ แต่ละตัวจำแนกจะมีกระบวนการทำงานของตัวเอง แล้วนำผลลัพธ์เหล่านั้นมาผ่านวิธีการรวบรวม (Combination หรือ Vote) และสุดท้ายนำไปตัดสินเพื่อเพิ่มประสิทธิภาพความแม่นยำให้กับแบบจำลองการเรียนรู้ (อัศวิน และวรภัทร, 2564) โดยจะใช้เทคนิคนี้กับการเรียนรู้ของเครื่องทั้งหมด 3 วิธี ได้แก่ วิธีนาอีฟเบส์ วิธีซัพพอร์ตเวกเตอร์แมชชีน และวิธีเพื่อนบ้านใกล้สุด k ตัว การเรียนรู้เชิงลึกทั้งหมด 3 วิธี ได้แก่ วิธีโครงข่ายประสาทเทียม วิธีโครงข่ายแบบวนซ้ำ และวิธีหน่วยความจำระยะสั้นยาว คณะผู้วิจัยได้เลือกใช้วิธีการโหวตเสียงข้างมาก (Majority Vote) ซึ่งวิธีนี้ง่ายต่อการใช้งาน (เดช และพยุง, 2554) โดยใช้แบบจำลอง VotingClassifier ในชุดคำสั่ง sklearn

3.9 การประเมินประสิทธิภาพของแบบจำลอง

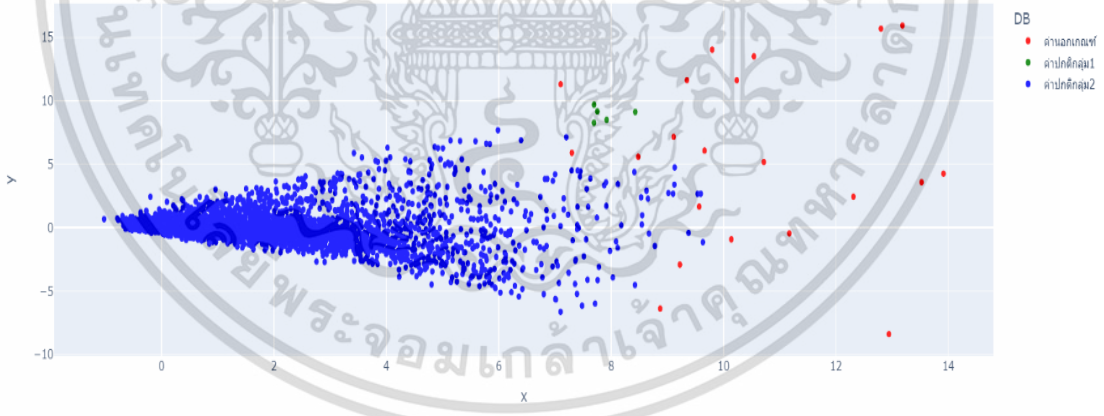
ประเมินประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก โดยพิจารณาจากค่าความแม่น ค่าความเที่ยง ค่าเรียกคืน และค่าประสิทธิภาพโดยรวม จากการใช้ชุดข้อมูลทดสอบจากการทำการกำจัดค่านอกเกณฑ์และไม่ได้กำจัดค่านอกเกณฑ์มาเป็นข้อมูลเข้า โดยใช้ชุดคำสั่ง sklearn

บทที่ 4
ผลการวิจัยและอภิปรายผล

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์ มีการนำข้อมูลมาจากเว็บไซต์ kaggle.com และทำการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF และนำมาสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีนาอีฟเบส์ วิธีซัพพอร์ตเวกเตอร์ แมชชีน และวิธีเพื่อนบ้านใกล้สุด k ตัว ใช้เทคนิคการเรียนรู้เชิงลึกด้วยวิธีโครงข่ายประสาทเทียม วิธีโครงข่ายแบบวนซ้ำ และวิธีหน่วยความจำระยะสั้นยาว นำค่าที่ได้มาเข้าวิธีการเรียนรู้แบบรวมกลุ่มในการโหวตวิธีการเรียนรู้ของเครื่องและวิธีการเรียนรู้เชิงลึก ด้วยภาษาไพทอน (Python) และนำมาเปรียบเทียบประสิทธิภาพในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์ดูจากค่าความแม่น ค่าความเที่ยง ค่าเรียกคืน และค่าประสิทธิภาพโดยผลการวิเคราะห์ได้นำเสนอตามลำดับดังนี้

4.1 ผลการวิเคราะห์การกำจัดค่านอกเกณฑ์

4.1.1 การกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN



รูปที่ 4.1 ผลการตรวจจับค่านอกเกณฑ์ด้วยวิธี DBSCAN

จากรูปที่ 4.1 เป็นผลจากการตรวจจับค่านอกเกณฑ์ด้วยวิธี DBSCAN แบ่งเป็น 3 กลุ่ม โดยแต่ละกลุ่มจะแสดงถึงค่าปกติและค่านอกเกณฑ์ ประกอบด้วย

- กลุ่มจุดสีแดง (-1) คือ ค่านอกเกณฑ์ ทั้งหมด 20 คำ
- กลุ่มจุดสีเขียว (1) คือ ค่าปกติกลุ่ม 1 ทั้งหมด 5 คำ
- กลุ่มจุดสีน้ำเงิน (0) คือ ค่าปกติกลุ่ม 2 ทั้งหมด 45,062 คำ

ตารางที่ 4.1 ผลการตรวจจับค่านอกเกณฑ์ด้วยวิธี DBSCAN

x	y	คำศัพท์	ค่ากลุ่ม DBSCAN
9.04	6.85	meter	-1
12.35	2.71	xls	-1
9.65	5.91	hpl	-1
8.32	4.2	nom	0
7.27	5.92	daren	-1
7.31	4.73	thanks	0
9.58	11.89	corp	-1
7.66	8.21	forwarded	1

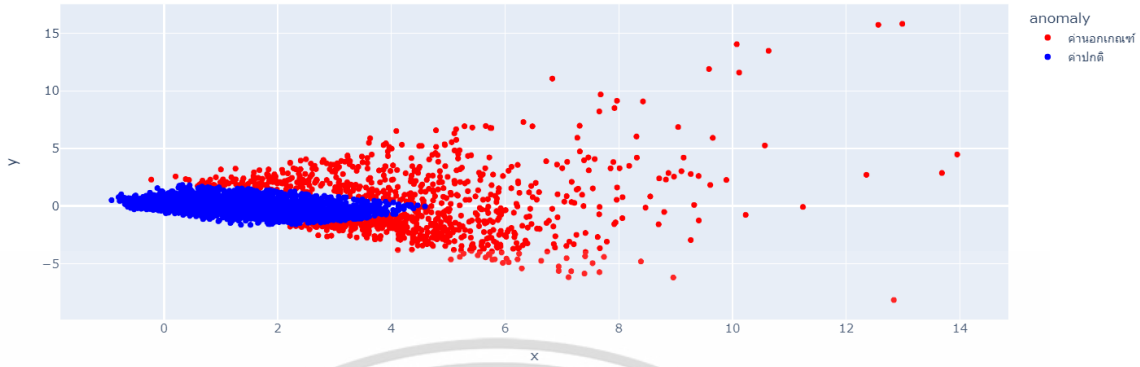
จากตารางที่ 4.1 เป็นตารางผลการตรวจจับค่านอกเกณฑ์ด้วยวิธี DBSCAN โดยค่า x และ y คือค่าเวกเตอร์ของคำ ค่ากลุ่ม DBSCAN คือ -1 หมายถึง ค่านอกเกณฑ์ 1 หมายถึง ค่าปกติกลุ่ม 1 และ 0 หมายถึง ค่าปกติกลุ่ม 2

ตารางที่ 4.2 ข้อความหลังจากกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN

ข้อความเดิม	ข้อความที่กำจัดค่านอกเกณฑ์แล้ว
enron methanol meter follow note gave monday preliminary flow data provided daren please override pop daily volume presently zero reflect daily activity obtain gas control change needed asap economics purposes	enron methanol follow note gave monday preliminary flow data provided please override pop daily volume presently zero reflect daily activity obtain gas control change needed asap economics purposes
hpl nom january see attached file hplnol xls hplnol xls	nom january see attached file hplnol hplnol

จากตารางที่ 4.2 เป็นตารางแสดงข้อความหลังจากกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN โดยนำคำที่เป็นค่านอกเกณฑ์จากวิธี DBSCAN ไปกำจัดออกจากข้อความเดิม

4.1.2 กำจัดค่านอกเกณฑ์ด้วยวิธีป่าไม้โดดเดี่ยว (Isolation Forest)



รูปที่ 4.2 ผลการตรวจจับค่านอกเกณฑ์ด้วยวิธีป่าไม้โดดเดี่ยว

จากรูป 4.2 เป็นรูปกผลจากการตรวจจับค่านอกเกณฑ์ด้วยวิธีป่าไม้โดดเดี่ยวแบ่งเป็น 2 กลุ่ม โดยแต่ละกลุ่มแสดงถึงค่าปกติและค่านอกเกณฑ์ ประกอบด้วย

- กลุ่มจุดสีแดง (-1) คือ ค่านอกเกณฑ์ ทั้งหมด 902 คำ
- กลุ่มจุดสีน้ำเงิน (1) คือ ค่าปกติ ทั้งหมด 44,185 คำ

ตารางที่ 4.3 ผลการตรวจจับค่านอกเกณฑ์ด้วยวิธีป่าไม้โดดเดี่ยว

x	y	คำศัพท์	คะแนนความผิดปกติ	ค่ากลุ่ม IF
7.68	9.68	enron	-0.13	-1
3.16	1.79	methanol	-0.04	-1
9.65	5.91	hpl	-0.13	-1
3.87	-1.38	follow	-0.02	-1
9.14	4.20	gas	-0.13	-1
7.31	4.73	thanks	-0.12	-1
8.32	4.2	nom	-0.13	-1
12.35	2.71	xls	-0.12	-1
1.80	-0.59	trip	0.08	1

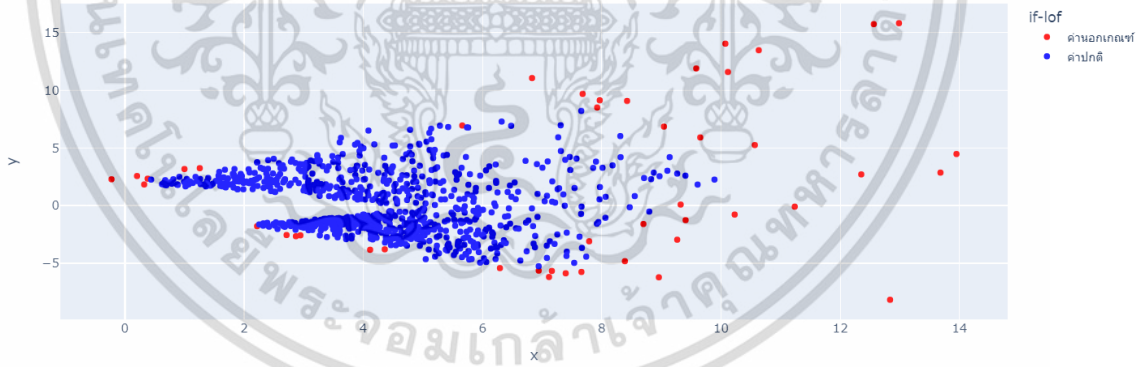
จากตารางที่ 4.3 เป็นตารางผลการตรวจจับค่านอกเกณฑ์ด้วยวิธีป่าไม้โดดเดี่ยวโดยค่า x และ y คือค่าเวกเตอร์ของคำ คะแนนความผิดปกติ คือ ค่าที่ใช้ในการแยกประเภทของข้อมูล ซึ่งคะแนนความความผิดปกตินั้นจะมีค่าตั้งแต่ 0 ถึง 1 โดยค่าเข้าใกล้ -1 จะหมายถึงข้อมูลที่มีแนวโน้มผิดปกติ ส่วนข้อมูลที่มีค่าเข้าใกล้ 1 จะถือว่าเป็นข้อมูลทั่วไปที่ไม่มีความผิดปกติ ค่ากลุ่ม IF คือ -1 หมายถึง ค่านอกเกณฑ์ และ 1 หมายถึง ค่าปกติ

ตารางที่ 4.4 ข้อความหลังจากกำจัดค่านอกเกณฑ์ด้วยวิธีป่าไม้โดดเดี่ยว

ข้อความเดิม	ข้อความที่กำจัดค่านอกเกณฑ์แล้ว
enron methanol meter follow note gave monday preliminary flow data provided daren please override pop daily volume presently zero reflect daily activity obtain gas control change needed asap economics purposes	gave preliminary override presently zero reflect obtain control asap economics purposes
hpl nom january see attached file hplnol xls hplnol xls	hplnol hplnol

จากตารางที่ 4.4 เป็นตารางแสดงข้อความหลังจากกำจัดค่านอกเกณฑ์ด้วยวิธีป่าไม้โดดเดี่ยว โดยนำคำที่เป็นค่านอกเกณฑ์จากวิธีป่าไม้โดดเดี่ยวไปกำจัดออกจากข้อความเดิม

4.1.3 กำจัดค่านอกเกณฑ์ด้วยวิธี IF-LOF



รูปที่ 4.3 ผลการตรวจจับค่านอกเกณฑ์ด้วยวิธี IF-LOF

จากรูป 4.3 เป็นรูปผลจากการตรวจจับค่านอกเกณฑ์ด้วยวิธี IF-LOF แบ่งเป็น 2 กลุ่ม โดยแต่ละกลุ่มแสดงถึงค่าปกติและค่านอกเกณฑ์ ประกอบด้วย

- กลุ่มจุดสีแดง (-1) คือ ค่านอกเกณฑ์ ทั้งหมด 46 คำ
- กลุ่มจุดสีน้ำเงิน (1) คือ ค่าปกติ ทั้งหมด 856 คำ

ตารางที่ 4.5 ผลลัพธ์การตรวจจับค่านอกเกณฑ์ด้วยวิธี IF-LOF

x	y	คำศัพท์	คะแนนความผิดปกติ	ค่ากลุ่ม IF-LOF
7.68	9.68	enron	-0.13	-1
9.04	6.85	meter	-0.14	-1
9.65	5.91	hpl	-0.13	-1
6.22	3.56	e	-0.11	1
12.35	2.71	xls	-0.12	-1
7.31	4.73	thanks	-0.13	1
9.58	11.89	corp	-0.14	-1
7.66	8.21	forwarded	-0.13	1

จากตารางที่ 4.5 เป็นตารางตัวอย่างผลลัพธ์การตรวจจับค่านอกเกณฑ์ด้วยวิธี IF-LOF โดยค่า x และ y คือค่าเวกเตอร์ของคำ คะแนนความผิดปกติ คือ ค่าที่ใช้ในการแยกประเภทของข้อมูล ซึ่งคะแนนความความผิดปกตินั้นจะมีค่าตั้งแต่ 0 ถึง 1 โดยค่าเข้าใกล้ -1 จะหมายถึงข้อมูลที่มีแนวโน้มผิดปกติส่วนข้อมูลที่มีค่าเข้าใกล้ 1 จะถือว่าเป็นข้อมูลทั่วไปที่ไม่มีความผิดปกติ ค่ากลุ่ม IF-LOF คือ -1 หมายถึง ค่านอกเกณฑ์ และ 1 หมายถึง ค่าปกติกลุ่ม

ตารางที่ 4.6 ข้อความหลังจากกำจัดค่านอกเกณฑ์ด้วยวิธี IF-LOF

ข้อความเดิม	ข้อความที่กำจัดค่านอกเกณฑ์แล้ว
enron methanol meter follow note gave monday preliminary flow data provided daren please override pop daily volume presently zero reflect daily activity obtain gas control change needed asap economics purposes	methanol follow note gave monday preliminary flow data provided daren please override pop daily volume presently zero reflect daily activity obtain gas control change needed asap economics purposes
hpl nom january see attached file hplnol xls hplnol xls	nom january see attached file hplnol hplnol

จากตารางที่ 4.6 เป็นตารางแสดงข้อความหลังจากกำจัดค่านอกเกณฑ์ด้วยวิธี IF-LOF โดยนำคำที่เป็นค่านอกเกณฑ์จากวิธี IF-LOF ไปกำจัดออกจากข้อความเดิม

4.2 ผลการวิเคราะห์การเรียนรู้ของเครื่อง

4.2.1 วิธีนาอีฟเบส์ (Naive Bayes)

จากการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีนาอีฟเบส์ในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์โดยใช้ชุดข้อมูลฝึกสอนที่ไม่มีการกำจัดค่านอกเกณฑ์มาเป็นข้อมูลเข้า และได้ใช้ชุดข้อมูลทดสอบในการประเมินประสิทธิภาพ ซึ่งได้ค่าความแม่น ค่าความเที่ยง ค่าเรียกคืน และค่าประสิทธิภาพโดยรวม ดังตารางที่ 4.7

ตารางที่ 4.7 ประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีนาอีฟเบส์ที่ไม่ได้ทำการกำจัดค่านอกเกณฑ์ โดยใช้ชุดข้อมูลทดสอบ

การเรียนรู้ของเครื่อง	ค่าความแม่น	ค่าความเที่ยง	ค่าเรียกคืน	ค่าประสิทธิภาพโดยรวม	เวลา (หน่วย : วินาที)
วิธีนาอีฟเบส์	98.07%	97.3%	96.00%	96.64%	0.7826

จากตารางที่ 4.7 แสดงประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีนาอีฟเบส์จากการใช้ชุดข้อมูลทดสอบที่ไม่มีการกำจัดค่านอกเกณฑ์มาใช้เป็นข้อมูลเข้า พบว่าประสิทธิภาพการทำนายของแบบจำลองนาอีฟเบส์มีค่าความแม่นเท่ากับ 98.07% ค่าความเที่ยงเท่ากับ 97.3% ค่าเรียกคืนเท่ากับ 96% และค่าประสิทธิภาพโดยรวมเท่ากับ 96.64%

จากนั้นทำการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีนาอีฟเบส์ในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์โดยใช้ชุดข้อมูลทดสอบที่กำจัดค่านอกเกณฑ์ 3 วิธี ได้แก่ วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF แล้วนำมาเป็นข้อมูลเข้า และได้ใช้ชุดข้อมูลทดสอบในการประเมินประสิทธิภาพ ซึ่งได้ค่าความแม่น ค่าความเที่ยง ค่าเรียกคืนและค่าประสิทธิภาพโดยรวม ดังตารางที่ 4.8

ตารางที่ 4.8 ประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีนาอีฟเบส์ที่ผ่านการกำจัดค่านอกเกณฑ์ โดยใช้ชุดข้อมูลทดสอบ

วิธีกำจัดค่านอกเกณฑ์	ค่าความแม่น	ค่าความเที่ยง	ค่าเรียกคืน	ค่าประสิทธิภาพโดยรวม	เวลา (หน่วย : วินาที)
DBSCAN	97.55%	97.47%	94.00%	95.70%	0.4847
ป่าไม้โดดเดี่ยว	95.36%	98.96%	84.89%	91.39%	0.2593
IF-LOF	97.36%	97.23%	93.56%	95.36%	1.0221

จากตารางที่ 4.8 แสดงประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีนาอีฟเบส์จากการใช้ชุดข้อมูลทดสอบที่กำจัดค่านอกเกณฑ์ 3 วิธี ได้แก่วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF พบว่าประสิทธิภาพการทำนายของแบบจำลองนาอีฟเบส์จากการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN มีค่าความแม่น ค่าเรียกคืน ค่าประสิทธิภาพโดยรวมมากที่สุด มีค่าเท่ากับ 97.55%, 94% และ 95.70% ตามลำดับ และการกำจัดค่านอกเกณฑ์ด้วยวิธีป่าไม้โดดเดี่ยวมีค่าความเที่ยงมากที่สุด มีค่าเท่ากับ 98.96% จึงสรุปได้ว่าการเรียนรู้ของเครื่องด้วยวิธีนาอีฟเบส์จากการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN ให้ค่าประสิทธิภาพการทำนายดีที่สุด

ตารางที่ 4.9 ประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีนาอีฟเบส์ที่ไม่ได้ทำการกำจัดค่านอกเกณฑ์และที่ผ่านการกำจัดค่านอกเกณฑ์ โดยใช้ชุดข้อมูลทดสอบ

วิธีกำจัดค่านอกเกณฑ์	ค่าความแม่น	ค่าความเที่ยง	ค่าเรียกคืน	ค่าประสิทธิภาพโดยรวม	เวลา (หน่วย : วินาที)
ไม่กำจัดค่านอกเกณฑ์	98.07%	97.30%	96.00%	96.64%	0.7826
DBSCAN	97.55%	97.47%	94.00%	95.70%	0.4847
ป่าไม้โดดเดี่ยว	95.36%	98.96%	84.89%	91.39%	0.2593
IF-LOF	97.36%	97.23%	93.56%	95.36%	1.0221

จากตารางที่ 4.9 แสดงประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีนาอีฟเบส์โดยใช้ชุดข้อมูลทดสอบจากชุดข้อมูลที่ไม่ทำการกำจัดค่านอกเกณฑ์เทียบกับชุดข้อมูลที่ทำการกำจัดค่านอกเกณฑ์ 3 วิธี ได้แก่วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF พบว่าประสิทธิภาพการทำนายของแบบจำลองนาอีฟเบส์จากชุดข้อมูลไม่ทำการกำจัดค่านอกเกณฑ์มีค่าความแม่น ค่าเรียกคืน ค่าประสิทธิภาพโดยรวมมากที่สุด มีค่าเท่ากับ 98.07%, 96.00% และ 96.64% ตามลำดับ และการกำจัดค่านอกเกณฑ์ด้วยวิธีป่าไม้โดดเดี่ยวมีค่าความเที่ยงมากที่สุด มีค่าเท่ากับ 98.96% จึงสรุปได้ว่าการเรียนรู้ของเครื่องด้วยวิธีนาอีฟเบส์จากชุดข้อมูลไม่ทำการกำจัดค่านอกเกณฑ์ให้ค่าประสิทธิภาพการทำนายดีที่สุด

4.2.2 วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

จากการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์โดยฟังก์ชันเคอร์เนลที่นำมาใช้คือฟังก์ชันเกาส์เซียนเรเดียลเบสิส (Gaussian Radial Basis Function-RBF) โดยใช้ชุดข้อมูลฝึกสอนที่ไม่มีการกำจัดค่านอกเกณฑ์มาใช้เป็นข้อมูลเข้า และได้ใช้ชุดข้อมูลทดสอบในการประเมินประสิทธิภาพ ซึ่งได้ค่าความแม่น ค่าความเที่ยง ค่าเรียกคืนและค่าประสิทธิภาพโดยรวม ดังตารางที่ 4.10

ตารางที่ 4.10 ประสิทธิภาพการทำนายของแบบจำลองวิธีซัพพอร์ตเวกเตอร์แมชชีนที่ไม่ได้ทำการกำจัดค่านอกเกณฑ์ โดยใช้ชุดข้อมูลทดสอบ

การเรียนรู้ของเครื่อง	ค่าความแม่น	ค่าความเที่ยง	ค่าเรียกคืน	ค่าประสิทธิภาพโดยรวม	เวลา (หน่วย : วินาที)
วิธีซัพพอร์ตเวกเตอร์แมชชีน	96.97%	91.72%	98.44%	94.96%	338.1371

จากตารางที่ 4.10 แสดงประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีซัพพอร์ตเวกเตอร์ แมชชีนจากชุดข้อมูลทดสอบที่ไม่มีการกำจัดค่านอกเกณฑ์มาใช้เป็นข้อมูลเข้า พบว่าประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนมีค่าความแม่นเท่ากับ 96.97% ค่าความเที่ยงเท่ากับ 91.72% ค่าเรียกคืนเท่ากับ 98.44% และค่าประสิทธิภาพโดยรวมเท่ากับ 94.96%

จากนั้นทำการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์โดยฟังก์ชันเคอร์เนลที่นำมาใช้คือฟังก์ชันเกาส์เซียนเรเดียลเบสิส โดยใช้ชุดข้อมูลฝึกสอนที่กำจัดค่านอกเกณฑ์ 3 วิธี ได้แก่วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF แล้วนำมาเป็นข้อมูลเข้า และได้ใช้ชุดข้อมูลทดสอบในการประเมินประสิทธิภาพ ซึ่งได้ค่าความแม่น ค่าความเที่ยง ค่าเรียกคืนและค่าประสิทธิภาพโดยรวม ดังตารางที่ 4.11

ตารางที่ 4.11 ประสิทธิภาพการทำนายของแบบจำลองวิธีซัพพอร์ตเวกเตอร์แมชชีนที่ผ่านการกำจัดค่านอกเกณฑ์ โดยใช้ชุดข้อมูลทดสอบ

วิธีกำจัดค่านอกเกณฑ์	ค่าความแม่น	ค่าความเที่ยง	ค่าเรียกคืน	ค่าประสิทธิภาพโดยรวม	เวลา (หน่วย : วินาที)
DBSCAN	**97.29%**	93.97%	**96.89%**	**95.40%**	3.7863
ป่าไม้โดดเดี่ยว	82.86%	82.86%	51.56%	63.56%	2.2413
IF-LOF	96.65%	**94.82%**	93.56%	94.18%	3.1778

จากตารางที่ 4.11 แสดงประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีซัพพอร์ตเวกเตอร์ แมชชีนจากการกำจัดค่านอกเกณฑ์ 3 วิธี ได้แก่วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF พบว่าประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนจากการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN มีค่าความแม่น ค่าเรียกคืน ค่าประสิทธิภาพโดยรวมมากที่สุด มีค่าเท่ากับ 97.29%, 96.89% และ 95.40% ตามลำดับ และการกำจัดค่านอกเกณฑ์ด้วยวิธี IF-LOF มีค่าความเที่ยงมากที่สุด มีค่าเท่ากับ 98.96% จึงสรุปได้ว่าการเรียนรู้ของเครื่องด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนจากการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN ให้ค่าประสิทธิภาพการทำนายดีที่สุด

ตารางที่ 4.12 ประสิทธิภาพการทำนายของแบบจำลองวิธีซัพพอร์ตเวกเตอร์แมชชีนไม่ได้ทำการกำจัดค่านอกเกณฑ์และที่ผ่านการกำจัดค่านอกเกณฑ์ โดยใช้ชุดข้อมูลทดสอบ

วิธีกำจัดค่านอกเกณฑ์	ค่าความแม่น	ค่าความเที่ยง	ค่าเรียกคืน	ค่าประสิทธิภาพโดยรวม	เวลา (หน่วย : วินาที)
ไม่กำจัดค่านอกเกณฑ์	96.97%	91.72%	**98.44%**	94.96%	338.1371
DBSCAN	**97.29%**	93.97%	96.89%	**95.40%**	3.7863
ป่าไม้โดดเดี่ยว	82.86%	82.86%	51.56%	63.56%	2.2413
IF-LOF	96.65%	**94.82%**	93.56%	94.18%	3.1778

จากตารางที่ 4.12 แสดงประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้ชุดข้อมูลทดสอบจากชุดข้อมูลที่ไม่ทำการกำจัดค่านอกเกณฑ์เทียบกับชุดข้อมูลที่ทำการกำจัดค่านอกเกณฑ์ 3 วิธี ได้แก่วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF พบว่าประสิทธิภาพการทำนายของแบบจำลองวิธีซัพพอร์ตเวกเตอร์แมชชีนจากชุดข้อมูลที่ทำการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN มีค่าความแม่นและค่าประสิทธิภาพโดยรวมมากที่สุด มีค่าเท่ากับ 97.29% และ 95.40% ตามลำดับ การกำจัดค่านอกเกณฑ์ด้วยวิธี IF-LOF มีค่าความเที่ยงมากที่สุด มีค่าเท่ากับ 94.82% และชุดข้อมูลที่ไม่ทำการกำจัดค่านอกเกณฑ์มีค่าเรียกคืนมากที่สุด มีค่าเท่ากับ 98.44% จึงสรุปได้ว่าการเรียนรู้ของเครื่องด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนจากชุดข้อมูลที่ทำการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN ให้ค่าประสิทธิภาพการทำนายดีที่สุด

4.2.3 วิธีเพื่อนบ้านใกล้สุด k ตัว (K-Nearest Neighbors)

จากการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีเพื่อนบ้านใกล้สุด k ตัว ในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์ โดยใช้วิธีระยะห่างยุคลิด ทำการพิจารณาค่า k ตั้งแต่ 1 ถึง 15 โดยใช้ชุดข้อมูลฝึกสอนที่ไม่มีการกำจัดค่านอกเกณฑ์มาใช้เป็นข้อมูลเข้า และได้ใช้ชุดข้อมูลทดสอบในการประเมินประสิทธิภาพ ซึ่งได้ค่าความแม่น ดังตารางที่ 4.13

ตารางที่ 4.13 ค่าความแม่นโดยใช้ค่า k เท่ากับ 1 ถึง 15 โดยใช้ข้อมูลที่ไม่มีการกำจัดค่านอกเกณฑ์โดยใช้ชุดข้อมูลทดสอบ

ค่า k	ค่าความแม่น
1	87.69%
2	83.05%
3	86.86%
4	84.54%
5	81.25%
6	81.31%
7	78.54%
8	78.67%
9	76.93%
10	77.45%
11	75.52%
12	75.84%
13	74.81%
14	75.00%
15	74.29%

จากตารางที่ 4.13 แสดงค่าความแม่นโดยใช้ค่า k เท่ากับ 1 ถึง 15 พบว่าที่ k เท่ากับ 1 มีค่าความแม่นมากที่สุด คิดเป็น 87.69% จึงใช้ค่า k เท่ากับ 1 ในการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้ชุดข้อมูลฝึกสอนที่ไม่มีการกำจัดค่านอกเกณฑ์มาเป็นข้อมูลเข้า และได้ใช้ชุดข้อมูลทดสอบในการประเมินประสิทธิภาพ ซึ่งได้ความแม่น ค่าความเที่ยง ค่าเรียกคืน และค่าประสิทธิภาพโดยรวม ดังตารางที่ 4.14

ตารางที่ 4.14 ประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีเพื่อนบ้านใกล้สุด k ตัว ที่ไม่ได้ทำการกำจัดค่านอกเกณฑ์

การเรียนรู้ของเครื่อง	ค่าความแม่น	ค่าความเที่ยง	ค่าเรียกคืน	ค่าประสิทธิภาพโดยรวม	เวลา (หน่วย : วินาที)
เพื่อนบ้านใกล้สุด k ตัว	87.69%	52.64%	97.56%	68.38%	0.5855

จากตารางที่ 4.14 แสดงประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีเพื่อนบ้านใกล้สุด k ตัว จากการใช้ชุดข้อมูลฝึกสอนที่ไม่มีการกำจัดค่านอกเกณฑ์มาใช้เป็นข้อมูลเข้า พบว่าประสิทธิภาพการทำนายของแบบจำลองวิธีเพื่อนบ้านใกล้สุด k ตัว มีค่าความแม่นเท่ากับ 87.69% ค่าความเที่ยงเท่ากับ 52.64% ค่าเรียกคืนเท่ากับ 97.56% และค่าประสิทธิภาพโดยรวมเท่ากับ 68.38%

จากนั้นทำการจากการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีเพื่อนบ้านใกล้สุด k ตัว ในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์ โดยใช้วิธีระยะห่างยุคลิด ทำการพิจารณาค่า k ตั้งแต่ 1 ถึง 15 โดยใช้ชุดข้อมูลฝึกสอนจากการกำจัดค่านอกเกณฑ์ 3 วิธี ได้แก่วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF แล้วนำมาเป็นข้อมูลเข้า และได้ใช้ชุดข้อมูลทดสอบในการประเมินประสิทธิภาพ ซึ่งได้ค่าความแม่น ดังตารางที่ 4.15

ตารางที่ 4.15 ค่าความแม่นโดยใช้ค่า k เท่ากับ 1 ถึง 15 โดยใช้ข้อมูลที่ผ่านการกำจัดค่านอกเกณฑ์โดยใช้ชุดข้อมูลทดสอบ

ค่า k	DBSCAN	ป่าไม้โดดเดี่ยว	IF-LOF
1	86.02%	83.12%	85.82%
2	82.86%	78.54%	82.93%
3	85.95%	78.61%	85.44%
4	82.41%	76.16%	81.83%
5	78.74%	76.55%	77.32%
6	79.12%	74.87%	77.77%
7	76.74%	75.06%	75.32%
8	77.45%	74.48%	75.84%
9	74.94%	74.48%	73.26%
10	75.45%	73.52%	73.52%
11	73.90%	73.65%	71.91%
12	74.36%	73.2%	72.23%
13	73.26%	73.26%	70.94%
14	73.58%	72.94%	71.39%
15	72.55%	73.00%	69.85%

จากตารางที่ 4.15 แสดงค่าความแม่นโดยใช้ค่า k เท่ากับ 1 ถึง 15 พบว่าที่ k เท่ากับ 1 จากการใช้ชุดข้อมูลฝึกสอนที่กำจัดค่านอกเกณฑ์ 3 วิธี ได้แก่วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF ได้ค่าความแม่นมากที่สุด คิดเป็น 86.02%, 83.12% และ 85.82% ตามลำดับ จึงใช้ค่า k เท่ากับ 1 ในการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีเพื่อนบ้านใกล้สุด k ตัว และได้ใช้ชุดข้อมูลทดสอบในการประเมินประสิทธิภาพ ซึ่งได้ความแม่น ค่าความเที่ยง ค่าเรียกคืน และค่าประสิทธิภาพโดยรวม ดังตารางที่ 4.16

ตารางที่ 4.16 ประสิทธิภาพการทำนายของแบบจำลองวิธีเพื่อนบ้านใกล้สุด k ตัว ที่ผ่านการกำจัดค่านอกเกณฑ์ โดยใช้ชุดข้อมูลทดสอบ

วิธีกำจัดค่านอกเกณฑ์	ค่าความแม่น	ค่าความเที่ยง	ค่าเรียกคืน	ค่าประสิทธิภาพโดยรวม	เวลา (หน่วย : วินาที)
DBSCAN	86.02%	50.41%	96.44%	66.21%	0.5008
ป่าไม้โดดเดี่ยว	83.12%	96.30%	5.78%	10.90%	0.2723
IF-LOF	85.82%	47.61%	95.33%	63.51%	0.4362

จากตารางที่ 4.16 แสดงประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีเพื่อนบ้านใกล้สุด k ตัว จากการใช้ชุดข้อมูลฝึกสอนที่กำจัดค่านอกเกณฑ์ 3 วิธี ได้แก่ วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF พบว่าประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีเพื่อนบ้านใกล้สุด k ตัว จากการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN มีค่าความแม่น ค่าเรียกคืน ค่าประสิทธิภาพโดยรวมมากที่สุด มีค่าเท่ากับ 86.02%, 96.44% และ 66.21% ตามลำดับ และการกำจัดค่านอกเกณฑ์ด้วยวิธีป่าไม้โดดเดี่ยว มีค่าความเที่ยงมากที่สุด มีค่าเท่ากับ 96.30% จึงสรุปได้ว่าการเรียนรู้ของเครื่องด้วยวิธีเพื่อนบ้านใกล้สุด k ตัว จากการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN ให้ค่าประสิทธิภาพการทำนายดีที่สุด

ตารางที่ 4.17 ประสิทธิภาพการทำนายของแบบจำลองวิธีเพื่อนบ้านใกล้สุด k ตัว ที่ไม่ได้ทำการกำจัดค่านอกเกณฑ์และที่ผ่านการกำจัดค่านอกเกณฑ์ โดยใช้ชุดข้อมูลทดสอบ

วิธีกำจัดค่านอกเกณฑ์	ค่าความแม่น	ค่าความเที่ยง	ค่าเรียกคืน	ค่าประสิทธิภาพโดยรวม	เวลา (หน่วย : วินาที)
ไม่กำจัดค่านอกเกณฑ์	87.69%	52.64%	97.56%	68.38%	0.5855
DBSCAN	86.02%	50.41%	96.44%	66.21%	0.5008
ป่าไม้โดดเดี่ยว	83.12%	96.30%	5.78%	10.90%	0.2723
IF-LOF	85.82%	47.61%	95.33%	63.51%	0.4362

จากตารางที่ 4.17 แสดงประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีเพื่อนบ้านใกล้สุด k ตัว โดยใช้ชุดข้อมูลทดสอบจากชุดข้อมูลที่ไม่ทำการกำจัดค่านอกเกณฑ์เทียบกับชุดข้อมูลที่ทำการกำจัดค่านอกเกณฑ์ 3 วิธี ได้แก่วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF พบว่าประสิทธิภาพการทำนายของแบบจำลองวิธีเพื่อนบ้านใกล้สุด k ตัว จากชุดข้อมูลไม่ทำการกำจัดค่านอกเกณฑ์มีค่าความแม่น ค่าเรียกคืน ค่าประสิทธิภาพโดยรวมมากที่สุด มีค่าเท่ากับ 87.69%, 97.56% และ 68.38% ตามลำดับ และการกำจัดค่านอกเกณฑ์ด้วยวิธีป่าไม้โดดเดี่ยวมีค่าความเที่ยงมากที่สุด มีค่า

เท่ากับ 96.30% จึงสรุปได้ว่าการเรียนรู้ของเครื่องด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวจากชุดข้อมูลไม่ทำการกำจัดค่านอกเกณฑ์ให้ค่าประสิทธิภาพการทำนายดีที่สุด

4.3 ผลการวิเคราะห์การเรียนรู้เชิงลึก

4.3.1 วิธีโครงข่ายประสาทเทียม (Artificial Neural Network)

จากการสร้างแบบจำลองการเรียนรู้เชิงลึกด้วยวิธีโครงข่ายประสาทเทียมในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์โดยใช้ชุดข้อมูลฝึกสอนที่ไม่มีการกำจัดค่านอกเกณฑ์มาเป็นข้อมูลเข้า และได้ใช้ชุดข้อมูลทดสอบในการประเมินประสิทธิภาพ ซึ่งได้ค่าความแม่น ค่าความเที่ยง ค่าเรียกคืนและค่าประสิทธิภาพโดยรวม ดังตารางที่ 4.18

ตารางที่ 4.18 ประสิทธิภาพการทำนายของแบบจำลองโครงข่ายประสาทเทียมที่ไม่ได้ทำการกำจัดค่านอกเกณฑ์ โดยใช้ชุดข้อมูลทดสอบ

การเรียนรู้เชิงลึก	ค่าความแม่น	ค่าความเที่ยง	ค่าเรียกคืน	ค่าประสิทธิภาพโดยรวม	เวลา (หน่วย : วินาที)
วิธีโครงข่ายประสาทเทียม	98.84%	96.95%	99.11%	98.02%	2303.0585

จากตารางที่ 4.18 แสดงประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีโครงข่ายประสาทเทียม โดยใช้ชุดข้อมูลทดสอบที่ไม่มีการกำจัดค่านอกเกณฑ์มาเป็นข้อมูลเข้า พบว่าประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีโครงข่ายประสาทเทียมมีค่าความแม่นเท่ากับ 98.84% ค่าความเที่ยงเท่ากับ 96.95% ค่าเรียกคืนเท่ากับ 99.11% และค่าประสิทธิภาพโดยรวมเท่ากับ 98.02%

จากนั้นทำการสร้างแบบจำลองการเรียนรู้เชิงลึกด้วยวิธีโครงข่ายประสาทเทียมในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์โดยใช้ชุดข้อมูลฝึกสอนที่กำจัดค่านอกเกณฑ์ 3 วิธี ได้แก่ วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF แล้วนำมาเป็นข้อมูลเข้า และได้ใช้ชุดข้อมูลทดสอบในการประเมินประสิทธิภาพ ซึ่งได้ได้ค่าความแม่น ค่าความเที่ยง ค่าเรียกคืนและค่าประสิทธิภาพโดยรวม ดังตารางที่ 4.19

ตารางที่ 4.19 ประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีโครงข่ายประสาทเทียมที่ผ่านการกำจัดค่านอกเกณฑ์ โดยใช้ชุดข้อมูลทดสอบ

วิธีกำจัดค่านอกเกณฑ์	ค่าความแม่น	ค่าความเที่ยง	ค่าเรียกคืน	ค่าประสิทธิภาพโดยรวม	เวลา (หน่วย : วินาที)
DBSCAN	**98.26%**	95.09%	**99.11%**	**97.06%**	1642.9794
ป่าไม้โดดเดี่ยว	95.49%	**97.74%**	86.44%	91.75%	1890.01977
IF-LOF	**98.26%**	97.10%	96.89%	96.99%	2243.0238

จากตารางที่ 4.19 แสดงประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีโครงข่ายประสาทเทียมจากชุดข้อมูลทดสอบที่กำจัดค่านอกเกณฑ์ 3 วิธี ได้แก่ วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF พบว่าประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีโครงข่ายประสาทเทียมจากการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN และวิธี IF-LOF มีค่าความแม่นมากที่สุด มีค่าเท่ากับ 98.26% นอกจากนี้วิธี DBSCAN ยังมีค่าเรียกคืนและค่าประสิทธิภาพโดยรวมมากที่สุด มีค่าเท่ากับ 99.11% และ 97.06% ตามลำดับ และการกำจัดค่านอกเกณฑ์ด้วยวิธีป่าไม้โดดเดี่ยวมีค่าความเที่ยงมากที่สุด มีค่าเท่ากับ 97.74% จึงสรุปได้ว่าการเรียนรู้เชิงลึกจากแบบจำลองด้วยวิธีโครงข่ายประสาทเทียมจากการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN ให้ค่าประสิทธิภาพการทำนายดีที่สุด

ตารางที่ 4.20 ประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีโครงข่ายประสาทเทียมที่ไม่ได้ทำการกำจัดค่านอกเกณฑ์และที่ผ่านการกำจัดค่านอกเกณฑ์ โดยใช้ชุดข้อมูลทดสอบ

วิธีกำจัดค่านอกเกณฑ์	ค่าความแม่น	ค่าความเที่ยง	ค่าเรียกคืน	ค่าประสิทธิภาพโดยรวม	เวลา (หน่วย : วินาที)
ไม่กำจัดค่านอกเกณฑ์	**98.84%**	96.95%	**99.11%**	**98.02%**	2303.0585
DBSCAN	98.26%	95.09%	**99.11%**	97.06%	1642.9794
ป่าไม้โดดเดี่ยว	95.49%	**97.74%**	86.44%	91.75%	1890.01977
IF-LOF	98.26%	97.10%	96.89%	96.99%	2243.0238

จากตารางที่ 4.20 แสดงประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีโครงข่ายประสาทเทียมโดยใช้ชุดข้อมูลทดสอบจากชุดข้อมูลที่ไม่ทำการกำจัดค่านอกเกณฑ์เทียบกับชุดข้อมูลที่ทำการกำจัดค่านอกเกณฑ์ 3 วิธี ได้แก่วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF พบว่าประสิทธิภาพการทำนายของแบบจำลองวิธีโครงข่ายประสาทเทียมจากชุดข้อมูลไม่ทำการกำจัดค่านอกเกณฑ์มีค่าความแม่น ค่าเรียกคืน ค่าประสิทธิภาพโดยรวมมากที่สุด มีค่าเท่ากับ 98.84%,

99.11% และ 98.02% ตามลำดับ และการกำจัดค่านอกเกณฑ์ด้วยวิธีป่าไม้โดดเดี่ยวมีค่าความเที่ยงมากที่สุด มีค่าเท่ากับ 97.74% จึงสรุปได้ว่าการเรียนรู้เชิงลึกด้วยวิธีโครงข่ายประสาทเทียมจากชุดข้อมูลไม่ทำการกำจัดค่านอกเกณฑ์ให้ค่าประสิทธิภาพการทำนายดีที่สุด

4.3.2 วิธีโครงข่ายประสาทแบบวนซ้ำ (Recurrent Neural Network)

จากการสร้างแบบจำลองการเรียนรู้เชิงลึกด้วยวิธีโครงข่ายประสาทแบบวนซ้ำในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์โดยใช้ชุดข้อมูลฝึกสอนที่ไม่มีการกำจัดค่านอกเกณฑ์มาเป็นข้อมูลเข้า และได้ใช้ชุดข้อมูลทดสอบในการประเมินประสิทธิภาพ ซึ่งได้ค่าความแม่น ค่าความเที่ยง ค่าเรียกคืนและค่าประสิทธิภาพโดยรวม ดังตารางที่ 4.21

ตารางที่ 4.21 ประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีโครงข่ายประสาทแบบวนซ้ำ โดยใช้ข้อมูลที่ไม่มีการกำจัดค่านอกเกณฑ์

การเรียนรู้เชิงลึก	ค่าความแม่น	ค่าความเที่ยง	ค่าเรียกคืน	ค่าประสิทธิภาพโดยรวม	เวลา (หน่วย : วินาที)
วิธีโครงข่ายประสาทแบบวนซ้ำ	97.68%	93.31%	99.11%	96.12%	11225.7590

จากตารางที่ 4.21 แสดงประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีโครงข่ายประสาทแบบวนซ้ำ โดยใช้ชุดข้อมูลทดสอบที่ไม่มีการกำจัดค่านอกเกณฑ์มาเป็นข้อมูลเข้า พบว่าประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีโครงข่ายประสาทแบบวนซ้ำมีค่าความแม่นเท่ากับ 97.68% ค่าความเที่ยงเท่ากับ 93.31% ค่าเรียกคืนเท่ากับ 99.11% และค่าประสิทธิภาพโดยรวมเท่ากับ 96.12%

จากนั้นทำการสร้างแบบจำลองการเรียนรู้เชิงลึกด้วยวิธีโครงข่ายประสาทแบบวนซ้ำในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์โดยชุดข้อมูลฝึกสอนที่กำจัดค่านอกเกณฑ์ 3 วิธี ได้แก่ วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF แลล้วนำมาเป็นข้อมูลเข้า และได้ใช้ชุดข้อมูลทดสอบในการประเมินประสิทธิภาพ ซึ่งได้ค่าความแม่น ค่าความเที่ยง ค่าเรียกคืนและค่าประสิทธิภาพโดยรวม ดังตารางที่ 4.22

ตารางที่ 4.22 ประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีโครงข่ายประสาทแบบวนซ้ำที่ผ่านการกำจัดค่านอกเกณฑ์ โดยใช้ชุดข้อมูลทดสอบ

วิธีกำจัดค่านอกเกณฑ์	ค่าความแม่น	ค่าความเที่ยง	ค่าเรียกคืน	ค่าประสิทธิภาพโดยรวม	เวลา (หน่วย : วินาที)
DBSCAN	**95.04%**	94.94%	**87.56%**	**91.09%**	8965.38801
ป่าไม้โดดเดี่ยว	87.18%	86.17%	66.44%	75.03%	8304.0357
IF-LOF	94.97%	**96.97%**	85.33%	90.78%	12503.9594

จากตารางที่ 4.22 แสดงประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีโครงข่ายประสาทแบบวนซ้ำจากชุดข้อมูลทดสอบที่กำจัดค่านอกเกณฑ์ 3 วิธี ได้แก่ วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF พบว่าประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีโครงข่ายประสาทแบบวนซ้ำจากชุดข้อมูลทดสอบที่กำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN มีค่าความแม่น ค่าเรียกคืน ค่าประสิทธิภาพโดยรวมมากที่สุด มีค่าเท่ากับ 95.04%, 87.56% และ 91.09% ตามลำดับ และการกำจัดค่านอกเกณฑ์ด้วยวิธี IF-LOF มีค่าความเที่ยงมากที่สุด มีค่าเท่ากับ 96.97% จึงสรุปได้ว่าการเรียนรู้เชิงลึกจากแบบจำลองด้วยวิธีโครงข่ายประสาทแบบวนซ้ำจากการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN ให้ค่าประสิทธิภาพการทำนายดีที่สุด

ตารางที่ 4.23 ประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีโครงข่ายประสาทแบบวนซ้ำที่ไม่ได้ทำการกำจัดค่านอกเกณฑ์และที่ผ่านการกำจัดค่านอกเกณฑ์ โดยใช้ชุดข้อมูลทดสอบ

วิธีกำจัดค่านอกเกณฑ์	ค่าความแม่น	ค่าความเที่ยง	ค่าเรียกคืน	ค่าประสิทธิภาพโดยรวม	เวลา (หน่วย : วินาที)
ไม่กำจัดค่านอกเกณฑ์	**97.68%**	93.31%	**99.11%**	**96.12%**	11225.7590
DBSCAN	95.04%	94.94%	87.56%	91.09%	8965.38801
ป่าไม้โดดเดี่ยว	87.18%	86.17%	66.44%	75.03%	8304.0357
IF-LOF	94.97%	**96.97%**	85.33%	90.78%	12503.9594

จากตารางที่ 4.23 แสดงประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีโครงข่ายประสาทแบบวนซ้ำโดยใช้ชุดข้อมูลทดสอบจากชุดข้อมูลที่ไม่ทำการกำจัดค่านอกเกณฑ์เทียบกับชุดข้อมูลที่ทำการกำจัดค่านอกเกณฑ์ 3 วิธี ได้แก่วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF พบว่าประสิทธิภาพการทำนายของแบบจำลองวิธีโครงข่ายประสาทแบบวนซ้ำจากชุดข้อมูลไม่ทำการกำจัดค่านอกเกณฑ์มีค่าความแม่น ค่าเรียกคืน ค่าประสิทธิภาพโดยรวมมากที่สุด มีค่าเท่ากับ 97.68%,

99.11% และ 96.12% ตามลำดับ และการกำจัดค่านอกเกณฑ์ด้วยวิธี IF-LOF มีค่าความเที่ยงมากที่สุด มีค่าเท่ากับ 96.97% จึงสรุปได้ว่าการเรียนรู้เชิงลึกด้วยวิธีโครงข่ายประสาทแบบวนซ้ำจากชุดข้อมูลไม่ทำการกำจัดค่านอกเกณฑ์ให้ค่าประสิทธิภาพการทำนายดีที่สุด

4.3.3 วิธีหน่วยความจำระยะสั้นยาว (Long Short-Term Memory)

จากการสร้างแบบจำลองการเรียนรู้เชิงลึกด้วยวิธีหน่วยความจำระยะสั้นยาวในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์โดยใช้ชุดข้อมูลฝึกสอนที่ไม่มีการกำจัดค่านอกเกณฑ์มาเป็นข้อมูลเข้า และได้ใช้ชุดข้อมูลทดสอบในการประเมินประสิทธิภาพ ซึ่งได้ค่าความแม่น ค่าความเที่ยง ค่าเรียกคืนและค่าประสิทธิภาพโดยรวม ดังตารางที่ 4.24

ตารางที่ 4.24 ประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีหน่วยความจำระยะสั้นยาวที่ไม่ได้ทำการกำจัดค่านอกเกณฑ์ โดยใช้ชุดข้อมูลทดสอบ

การเรียนรู้เชิงลึก	ค่าความแม่น	ค่าความเที่ยง	ค่าเรียกคืน	ค่าประสิทธิภาพโดยรวม	เวลา (หน่วย : วินาที)
วิธีหน่วยความจำระยะสั้นยาว	95.94%	97.78%	88.00%	92.63%	5004.9689

จากตารางที่ 4.24 แสดงประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีหน่วยความจำระยะสั้นยาว โดยใช้ชุดข้อมูลทดสอบที่ไม่มีการกำจัดค่านอกเกณฑ์มาเป็นข้อมูลเข้า พบว่าประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีหน่วยความจำระยะสั้นยาว มีค่าความแม่นเท่ากับ 95.94% ค่าความเที่ยงเท่ากับ 97.78% ค่าเรียกคืนเท่ากับ 88% และค่าประสิทธิภาพโดยรวมเท่ากับ 92.63%

จากนั้นทำการสร้างแบบจำลองการเรียนรู้เชิงลึกด้วยวิธีหน่วยความจำระยะสั้นยาวในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์โดยใช้ชุดข้อมูลฝึกสอนที่กำจัดค่านอกเกณฑ์ 3 วิธี ได้แก่วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF มาเป็นข้อมูลเข้า และได้ใช้ชุดข้อมูลทดสอบในการประเมินประสิทธิภาพ ซึ่งได้ค่าความแม่น ค่าความเที่ยง ค่าเรียกคืนและค่าประสิทธิภาพโดยรวม ดังตารางที่ 4.25

ตารางที่ 4.25 ประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีหน่วยความจำระยะสั้นยาวที่ผ่านการกำจัดค่านอกเกณฑ์ โดยใช้ชุดข้อมูลทดสอบ

วิธีกำจัดค่านอกเกณฑ์	ค่าความแม่น	ค่าความเที่ยง	ค่าเรียกคืน	ค่าประสิทธิภาพโดยรวม	เวลา (หน่วย : วินาที)
DBSCAN	97.81%	98.15%	94.22%	96.15%	3564.8449
ป่าไม้โดดเดี่ยว	96.13%	94.93%	91.55%	93.21%	4645.1194
IF-LOF	97.29%	98.11%	92.44%	95.19%	3684.4496

จากตารางที่ 4.25 แสดงประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีหน่วยความจำระยะสั้นยาวจากชุดข้อมูลฝึกสอนที่กำจัดค่านอกเกณฑ์ 3 วิธี ได้แก่ วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF พบว่าประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีหน่วยความจำระยะสั้นยาวจากการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN มีค่าความแม่น ค่าความเที่ยง ค่าเรียกคืน และค่าประสิทธิภาพโดยรวมมากที่สุด มีค่าเท่ากับ 97.81%, 98.15%, 94.22% และ 96.15% ตามลำดับ จึงสรุปได้ว่าการเรียนรู้เชิงลึกจากแบบจำลองด้วยวิธีหน่วยความจำระยะสั้นยาวจากการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN ให้ค่าประสิทธิภาพการทำนายดีที่สุด

ตารางที่ 4.26 ประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีหน่วยความจำระยะสั้นยาวที่ไม่ได้ทำการกำจัดค่านอกเกณฑ์และที่ผ่านการกำจัดค่านอกเกณฑ์ โดยใช้ชุดข้อมูลทดสอบ

วิธีกำจัดค่านอกเกณฑ์	ค่าความแม่น	ค่าความเที่ยง	ค่าเรียกคืน	ค่าประสิทธิภาพโดยรวม	เวลา (หน่วย : วินาที)
ไม่กำจัดค่านอกเกณฑ์	95.94%	97.78%	88.00%	92.63%	5004.9689
DBSCAN	97.81%	98.15%	94.22%	96.15%	3564.8449
ป่าไม้โดดเดี่ยว	96.13%	94.93%	91.55%	93.21%	4645.1194
IF-LOF	97.29%	98.11%	92.44%	95.19%	3684.4496

จากตารางที่ 4.26 แสดงประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีหน่วยความจำระยะสั้นยาวโดยใช้ชุดข้อมูลทดสอบจากชุดข้อมูลที่ไม่ทำการกำจัดค่านอกเกณฑ์เทียบกับชุดข้อมูลที่ทำการกำจัดค่านอกเกณฑ์ 3 วิธี ได้แก่วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF พบว่าประสิทธิภาพการทำนายของแบบจำลองวิธีหน่วยความจำระยะสั้นยาวจากชุดข้อมูลทำการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN มีค่าความแม่น ค่าความเที่ยง ค่าเรียกคืน ค่าประสิทธิภาพโดยรวมมากที่สุด มีค่าเท่ากับ 97.81%, 98.15%, 94.22% และ 96.15% ตามลำดับ จึงสรุปได้ว่าการเรียนรู้เชิง

ลึกด้วยวิธีหน่วยความจำระยะสั้นยาวจากชุดข้อมูลทำการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN ให้ค่าประสิทธิภาพการทำนายดีที่สุด

4.4 ผลการวิเคราะห์การเรียนรู้แบบรวมกลุ่ม (Ensemble Learning)

4.4.1 การเรียนรู้แบบรวมกลุ่มของการเรียนรู้ของเครื่อง

จากการสร้างแบบจำลองการเรียนรู้แบบรวมกลุ่มของการเรียนรู้ของเครื่อง คณะผู้วิจัยได้เลือกใช้วิธีการโหวตเสียงข้างมาก เพื่อเพิ่มประสิทธิภาพให้กับแบบจำลองการเรียนรู้ของเครื่อง โดยใช้ชุดข้อมูลฝึกสอนที่ไม่มีการกำจัดค่านอกเกณฑ์มาเป็นข้อมูลเข้า และได้ใช้ชุดข้อมูลทดสอบในการประเมินประสิทธิภาพ ซึ่งได้ค่าความแม่น ค่าความเที่ยง ค่าเรียกคืน และค่าประสิทธิภาพโดยรวม ดังตารางที่ 4.27

ตารางที่ 4.27 ประสิทธิภาพการทำนายของแบบจำลองการเรียนรู้แบบรวมกลุ่มของการเรียนรู้ของเครื่อง โดยใช้ข้อมูลที่ไม่มีการกำจัดค่านอกเกณฑ์ โดยใช้ชุดข้อมูลทดสอบ

การเรียนรู้แบบรวมกลุ่ม	ค่าความแม่น	ค่าความเที่ยง	ค่าเรียกคืน	ค่าประสิทธิภาพโดยรวม	เวลา (หน่วย : วินาที)
การเรียนรู้ของเครื่อง	97.68%	93.31%	99.11%	96.12%	3.6586

จากตารางที่ 4.27 แสดงประสิทธิภาพการทำนายของแบบจำลองการเรียนรู้แบบรวมกลุ่มโดยใช้ชุดข้อมูลฝึกสอนที่ไม่มีการกำจัดค่านอกเกณฑ์มาเป็นข้อมูลเข้า พบว่าประสิทธิภาพการทำนายของแบบจำลองการเรียนรู้แบบรวมกลุ่มมีค่าความแม่นเท่ากับ 97.68% ค่าความเที่ยงเท่ากับ 93.31% ค่าเรียกคืนเท่ากับ 99.11% และค่าประสิทธิภาพโดยรวมเท่ากับ 96.12%

จากนั้นสร้างแบบจำลองการเรียนรู้แบบรวมกลุ่มของการเรียนรู้ของเครื่องโดยใช้วิธีการโหวตเสียงข้างมาก เพื่อเพิ่มประสิทธิภาพความแม่นยำให้กับแบบจำลองการเรียนรู้ โดยใช้ชุดข้อมูลฝึกสอนที่กำจัดค่านอกเกณฑ์ 3 วิธี ได้แก่ วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF แล้วนำมาเป็นข้อมูลเข้า และได้ใช้ชุดข้อมูลทดสอบในการประเมินประสิทธิภาพ ซึ่งได้ค่าความแม่น ค่าความเที่ยง ค่าเรียกคืน และค่าประสิทธิภาพโดยรวม ดังตารางที่ 4.28

ตารางที่ 4.28 ประสิทธิภาพการทำนายของแบบจำลองการเรียนรู้แบบรวมกลุ่มของการเรียนรู้ของเครื่องที่ผ่านการกำจัดค่านอกเกณฑ์ โดยใช้ชุดข้อมูลทดสอบ

วิธีกำจัดค่านอกเกณฑ์	ค่าความแม่น	ค่าความเที่ยง	ค่าเรียกคืน	ค่าประสิทธิภาพโดยรวม	เวลา (หน่วย : วินาที)
DBSCAN	**98.13%**	95.46%	**98.22%**	**96.82%**	4.2339
ป่าไม้โดดเดี่ยว	88.60%	**96.91%**	62.67%	76.11%	3.4499
IF-LOF	98.00%	96.25%	96.89%	96.57%	3.7798

จากตารางที่ 4.28 แสดงประสิทธิภาพการทำนายของแบบจำลองการเรียนรู้แบบรวมกลุ่มจากชุดข้อมูลทดสอบที่กำจัดค่านอกเกณฑ์ 3 วิธี ได้แก่ วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF พบว่าประสิทธิภาพการทำนายของแบบจำลองการเรียนรู้แบบรวมกลุ่มจากการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN มีค่าความแม่น ค่าเรียกคืน ค่าประสิทธิภาพโดยรวมมากที่สุด มีค่าเท่ากับ 98.13%, 98.22% และ 96.82% ตามลำดับ และการกำจัดค่านอกเกณฑ์ด้วยวิธีป่าไม้โดดเดี่ยวมีค่าความเที่ยงมากที่สุด มีค่าเท่ากับ 96.91% จึงสรุปได้ว่าการเรียนรู้ของเครื่องจากการเรียนรู้แบบรวมกลุ่มการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN ให้ค่าประสิทธิภาพการทำนายดีที่สุด

ตารางที่ 4.29 ประสิทธิภาพการทำนายของแบบจำลองการเรียนรู้แบบรวมกลุ่มของการเรียนรู้ของเครื่องที่ไม่ได้ทำการกำจัดค่านอกเกณฑ์และที่ผ่านการกำจัดค่านอกเกณฑ์ โดยใช้ชุดข้อมูลทดสอบ

วิธีกำจัดค่านอกเกณฑ์	ค่าความแม่น	ค่าความเที่ยง	ค่าเรียกคืน	ค่าประสิทธิภาพโดยรวม	เวลา (หน่วย : วินาที)
ไม่กำจัดค่านอกเกณฑ์	97.68%	93.31%	**99.11%**	96.12%	3.6586
DBSCAN	**98.13%**	95.46%	98.22%	**96.82%**	4.2339
ป่าไม้โดดเดี่ยว	88.60%	**96.91%**	62.67%	76.11%	3.4499
IF-LOF	98.00%	96.25%	96.89%	96.57%	3.7798

จากตารางที่ 4.29 แสดงประสิทธิภาพการทำนายของแบบจำลองการเรียนรู้แบบรวมกลุ่มของการเรียนรู้ของเครื่องโดยใช้ชุดข้อมูลทดสอบจากชุดข้อมูลที่ไม่ทำการกำจัดค่านอกเกณฑ์เทียบกับชุดข้อมูลที่ทำการกำจัดค่านอกเกณฑ์ 3 วิธี ได้แก่วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF พบว่าการเรียนรู้ของเครื่องแบบจำลองการเรียนรู้แบบรวมกลุ่มโดยใช้ชุดข้อมูลทดสอบที่กำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN มีค่าความแม่นและค่าประสิทธิภาพโดยรวมมากที่สุด มีค่าเท่ากับ 98.13%, 96.82% ตามลำดับ กำจัดค่านอกเกณฑ์ด้วยวิธีป่าไม้โดดเดี่ยวมีค่าความเที่ยงมากที่สุด มีค่าเท่ากับ

96.91% และข้อมูลทดสอบที่ไม่ทำการกำจัดค่านอกเกณฑ์มีค่าเรียกคืนมากที่สุด มีค่าเท่ากับ 99.11% จึงสรุปได้ว่าประสิทธิภาพการทำนายของการเรียนรู้ของเครื่องแบบจำลองการเรียนรู้แบบรวมกลุ่มโดยใช้ชุดข้อมูลที่ทำการกำจัดค่านอกเกณฑ์ให้ผลที่ดีกว่า

4.4.2 การเรียนรู้แบบรวมกลุ่มของการเรียนรู้เชิงลึก

สร้างแบบจำลองการเรียนรู้แบบรวมกลุ่มของการเรียนรู้เชิงลึก คณะผู้วิจัยได้เลือกใช้วิธีการโหวตเสียงข้างมาก เพื่อเพิ่มประสิทธิภาพความแม่นยำให้กับแบบจำลองการเรียนรู้เชิงลึก โดยใช้ชุดข้อมูลฝึกสอนที่ไม่มีการกำจัดค่านอกเกณฑ์มาเป็นข้อมูลเข้า และใช้ชุดข้อมูลทดสอบในการประเมินประสิทธิภาพ ซึ่งได้ค่าความแม่น ค่าความเที่ยง ค่าเรียกคืน และค่าประสิทธิภาพโดยรวม ดังตารางที่ 4.30

ตารางที่ 4.30 ประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีการเรียนรู้แบบรวมกลุ่มของการเรียนรู้เชิงลึกที่ไม่มีการกำจัดค่านอกเกณฑ์ โดยใช้ชุดข้อมูลทดสอบ

การเรียนรู้แบบรวมกลุ่ม	**ค่าความแม่น**	**ค่าความเที่ยง**	**ค่าเรียกคืน**	**ค่าประสิทธิภาพโดยรวม**	**เวลา (หน่วย : วินาที)**
การเรียนรู้เชิงลึก	98.52%	97.55%	97.33%	97.44%	70.8726

จากตารางที่ 4.30 แสดงประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีการเรียนรู้แบบรวมกลุ่มของการเรียนรู้เชิงลึกโดยใช้ชุดข้อมูลทดสอบที่ไม่มีการกำจัดค่านอกเกณฑ์มาเป็นข้อมูลเข้า พบว่าประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีการเรียนรู้แบบรวมกลุ่มของการเรียนรู้เชิงลึกมีค่าความแม่นเท่ากับ 98.52% ค่าความเที่ยงเท่ากับ 97.55% ค่าเรียกคืนเท่ากับ 97.33% และค่าประสิทธิภาพโดยรวมเท่ากับ 97.44%

จากนั้นสร้างแบบจำลองด้วยวิธีการเรียนรู้แบบรวมกลุ่มของการเรียนรู้เชิงลึกโดยใช้วิธีการโหวตเสียงข้างมาก เพื่อเพิ่มประสิทธิภาพความแม่นยำให้กับแบบจำลองการเรียนรู้ โดยใช้ชุดข้อมูลฝึกสอนที่กำจัดค่านอกเกณฑ์ 3 วิธี ได้แก่ วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF แล้วนำมาเป็นข้อมูลเข้า ใช้ชุดข้อมูลทดสอบในการประเมินประสิทธิภาพ ซึ่ง ได้ค่าความแม่น ค่าความเที่ยง ค่าเรียกคืน และค่าประสิทธิภาพโดยรวม ดังตารางที่ 4.31

ตารางที่ 4.31 ประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีการเรียนรู้แบบรวมกลุ่มของการเรียนรู้เชิงลึกที่ผ่านการกำจัดค่านอกเกณฑ์ โดยใช้ชุดข้อมูลทดสอบ

วิธีกำจัดค่านอกเกณฑ์	ค่าความแม่น	ค่าความเที่ยง	ค่าเรียกคืน	ค่าประสิทธิภาพโดยรวม	เวลา (หน่วย : วินาที)
DBSCAN	98.26%	96.89%	97.11%	97.00%	59.2709
ป่าไม้โดดเดี่ยว	95.68%	97.05%	87.78%	92.18%	49.4138
IF-LOF	97.74%	97.48%	94.67%	96.05%	68.1231

จากตารางที่ 4.31 แสดงประสิทธิภาพการทำนายของแบบจำลองด้วยวิธีการเรียนรู้แบบรวมกลุ่มของการเรียนรู้เชิงลึก จากชุดข้อมูลทดสอบที่กำจัดค่านอกเกณฑ์ 3 วิธี ได้แก่ วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF พบว่าประสิทธิภาพการทำนายของแบบจำลองการเรียนรู้แบบรวมกลุ่มของการเรียนรู้เชิงลึกจากการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN มีค่าความแม่น ค่าเรียกคืน และค่าประสิทธิภาพโดยรวมมากที่สุด มีค่าเท่ากับ 98.26 %, 97.11% และ 97% ตามลำดับ และการกำจัดค่านอกเกณฑ์ด้วยวิธี IF-LOF มีค่าความเที่ยงมากที่สุด มีค่าเท่ากับ 97.48% จึงสรุปได้ว่าการเรียนรู้เชิงลึกจากการเรียนรู้แบบรวมกลุ่มจากการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN ให้ค่าประสิทธิภาพการทำนายดีที่สุด

ตารางที่ 4.32 ประสิทธิภาพการทำนายของแบบจำลองการเรียนรู้แบบรวมกลุ่มของการเรียนรู้ของเครื่องที่ไม่ได้ทำการกำจัดค่านอกเกณฑ์และที่ผ่านการกำจัดค่านอกเกณฑ์ โดยใช้ชุดข้อมูลทดสอบ

วิธีกำจัดค่านอกเกณฑ์	ค่าความแม่น	ค่าความเที่ยง	ค่าเรียกคืน	ค่าประสิทธิภาพโดยรวม	เวลา (หน่วย : วินาที)
ไม่กำจัดค่านอกเกณฑ์	98.52%	97.55%	97.33%	97.44%	70.8726
DBSCAN	98.26%	96.89%	97.11%	97.00%	59.2709
ป่าไม้โดดเดี่ยว	95.68%	97.05%	87.78%	92.18%	49.4138
IF-LOF	97.74%	97.48%	94.67%	96.05%	68.1231

จากตารางที่ 4.32 แสดงประสิทธิภาพการทำนายของแบบจำลองการเรียนรู้แบบรวมกลุ่มของการเรียนรู้เชิงลึกโดยใช้ชุดข้อมูลทดสอบจากชุดข้อมูลที่ไม่ทำการกำจัดค่านอกเกณฑ์เทียบกับชุดข้อมูลที่ทำการกำจัดค่านอกเกณฑ์ 3 วิธี ได้แก่วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF พบว่าการเรียนรู้เชิงลึกแบบจำลองการเรียนรู้แบบรวมกลุ่มโดยใช้ชุดข้อมูลทดสอบที่ไม่ทำการกำจัดค่านอกเกณฑ์มีค่าความแม่น ค่าความเที่ยง ค่าเรียกคืน และค่าประสิทธิภาพโดยรวมมากที่สุด มีค่าเท่ากับ

98.52%, 97.55%, 97.33% และ 97.44% ตามลำดับ จึงสรุปได้ว่าประสิทธิภาพการทำนายของการเรียนรู้เชิงลึกแบบจำลองการเรียนรู้แบบรวมกลุ่มโดยใช้ชุดข้อมูลที่ไม่ทำการกำจัดค่านอกเกณฑ์ให้ผลที่ดีกว่า แต่ข้อมูลที่ใช้จะมากขึ้นทำให้ใช้เวลานานกว่า ซึ่งอาจทำให้การวิเคราะห์อาจผิดพลาดได้ เนื่องจากยังมีค่านอกเกณฑ์อยู่ จึงใช้แบบจำลองการเรียนรู้แบบรวมกลุ่มของการเรียนรู้เชิงลึกโดยกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN มาใช้ในการสรุปผลต่อไป เนื่องจากมีค่าความแม่น ค่าเรียกคืน และค่าประสิทธิภาพโดยรวมมากที่สุด มีค่าเท่ากับ 98.26 %, 97.11% และ 97% ตามลำดับ

4.5 อภิปรายผลการวิจัย

การเปรียบเทียบจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์ โดยจะทำการกำจัดค่านอกเกณฑ์ทั้งหมด 3 วิธี คือ วิธี DBSCAN วิธีป่าไม้โดดเดี่ยวและวิธี IF-LOF ซึ่งพบว่าวิธี DBSCAN ให้ค่าประสิทธิภาพของแบบจำลองดีที่สุดในการกำจัดค่านอกเกณฑ์ของทั้งการเรียนรู้ของเครื่อง และการเรียนรู้เชิงลึกรองลงมาเป็นวิธี IF-LOF และวิธีป่าไม้โดดเดี่ยวโดยพิจารณาจากค่าความแม่น ค่าความเที่ยง ค่าเรียกคืนและค่าประสิทธิภาพโดยรวม ซึ่งแตกต่างกับงานวิจัยของ Hossain et al. (2021) กล่าวว่าวิธีการเรียนรู้ของเครื่องโดยใช้วิธีการตรวจจับค่านอกเกณฑ์ด้วยวิธี DBSCAN มีประสิทธิภาพมากที่สุดและในการเรียนรู้เชิงลึกโดยใช้วิธีการตรวจจับค่านอกเกณฑ์ด้วยวิธีป่าไม้โดดเดี่ยวมีประสิทธิภาพมากที่สุด เนื่องจากคณะผู้วิจัยใช้การเรียนรู้เชิงลึกการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN มีประสิทธิภาพมากที่สุด โดยในงานวิจัยของคณะผู้วิจัยมีการกำจัดค่านอกเกณฑ์ด้วยวิธีป่าไม้โดดเดี่ยวมีการตัดคำออกไปเป็นจำนวนมากทำให้ข้อมูลที่นำเข้าไปในแบบจำลองมีจำนวนน้อยลง เนื่องจากวิธี DBSCAN มีการตัดคำที่น้อยซึ่งมีการตัดคำไปทั้งหมด 20 คำ ทำให้ได้ค่าที่ดีกว่าวิธีป่าไม้โดดเดี่ยวซึ่งตัดคำไปทั้งหมด 902 คำ และวิธี IF-LOF ซึ่งมีการตัดคำไปทั้งหมด 46 คำ

ข้อมูลที่ทำการกำจัดค่านอกเกณฑ์แล้วจะนำไปเข้าแบบจำลองการเรียนรู้ของเครื่องทั้งหมด 3 วิธี คือ วิธีนาอีฟเบส์ วิธีซัพพอร์ตเวกเตอร์แมชชีน และวิธีเพื่อนบ้านใกล้สุด k ตัว ซึ่งพบว่าการเรียนรู้ของเครื่องโดยวัดประสิทธิภาพการทำนายจากค่าความแม่น ค่าความเที่ยง ค่าเรียกคืน และค่าประสิทธิภาพโดยรวม ผลการศึกษาพบว่าวิธีนาอีฟเบส์มีค่าความแม่น ค่าความเที่ยง ค่าเรียกคืนและค่าประสิทธิภาพโดยรวมสูงที่สุดด้วยการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN คิดเป็น 97.55% 97.47% 94.00% และ 95.70% ตามลำดับ ซึ่งสอดคล้องกับงานวิจัยของ Awad and Elseuofi (2011) ได้ศึกษาวิธีการเรียนรู้ของเครื่องทั้งหมด 6 วิธี ประกอบด้วยวิธีนาอีฟเบส์ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีโครงข่ายประสาทเทียม วิธีซัพพอร์ตเวกเตอร์แมชชีน ระบบภูมิคุ้มกันเทียมและทฤษฎีรัฟเซต โดยวัดประสิทธิภาพจากค่าความแม่น ค่าความเที่ยง และค่าการเรียกคืนซึ่งพบว่าวิธีนาอีฟเบส์มีค่าความแม่น ค่าความเที่ยง และค่าการเรียกคืนสูงที่สุด

ข้อมูลที่ทำการกำจัดค่านอกเกณฑ์แล้วจะนำไปเข้าแบบจำลองการเรียนรู้เชิงลึกทั้งหมด 3 วิธี คือ วิธีโครงข่ายประสาทเทียม วิธีโครงข่ายแบบวนซ้ำ และวิธีหน่วยความจำระยะสั้นยาว ผลการศึกษาพบว่าวิธีโครงข่ายประสาทเทียมมีค่าความแม่น ค่าความเที่ยง ค่าเรียกคืนและค่าประสิทธิภาพโดยรวม

มากที่สุดด้วยการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN คิดเป็น 98.26% 95.09% 99.11% และ 97.06% ตามลำดับ

ข้อมูลที่ทำการกำจัดค่านอกเกณฑ์แล้วจะใช้วิธีการเรียนรู้แบบรวมกลุ่มกับวิธีการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกเพื่อเพิ่มประสิทธิภาพความแม่นยำให้กับแบบจำลองการเรียนรู้ ผลการศึกษาพบว่าวิธีการเรียนรู้แบบรวมกลุ่มในการเรียนรู้ของเครื่องใช้การตรวจจับค่านอกเกณฑ์ด้วยวิธี DBSCAN ให้ผลที่ดีกว่า และการเรียนรู้แบบรวมกลุ่มช่วยเพิ่มประสิทธิภาพของแบบจำลอง ซึ่งมีค่าความแม่น ค่าความเที่ยง ค่าเรียกคืนและค่าประสิทธิภาพโดยรวม คิดเป็น 98.13% 95.46% 98.22% และ 96.82% ตามลำดับและวิธีการเรียนรู้แบบรวมกลุ่มในการเรียนรู้เชิงลึกใช้การตรวจจับค่านอกเกณฑ์ด้วยวิธี DBSCAN ให้ผลที่ดีกว่าและการเรียนรู้แบบรวมกลุ่มช่วยเพิ่มประสิทธิภาพของแบบจำลองมีค่าความแม่น ค่าความเที่ยง ค่าเรียกคืนและค่าประสิทธิภาพโดยรวม คิดเป็น 98.26% 96.89% 97.11% และ 97.00% ตามลำดับ

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

จากการศึกษาจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์ โดยใช้การกำจัดค่านอกเกณฑ์ แล้วนำเข้าแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก แล้วใช้การเรียนรู้แบบรวมกลุ่ม ในการเปรียบเทียบประสิทธิภาพของแบบจำลอง ซึ่งมีวัตถุประสงค์ดังนี้ 1) ศึกษาวิธีการตรวจจับค่านอกเกณฑ์ในประเภทของข้อมูลที่เป็นข้อความ 2) ศึกษาวิธีการเรียนรู้ของเครื่องและวิธีการเรียนรู้เชิงลึกในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์ 3) เปรียบเทียบวิธีการเรียนรู้ของเครื่องและวิธีการเรียนรู้เชิงลึก 4)เปรียบเทียบวิธีการเรียนรู้แบบรวมกลุ่มในวิธีการเรียนรู้ของเครื่องและวิธีการเรียนรู้เชิงลึก

การศึกษาครั้งนี้ได้นำข้อมูลมาจากเว็บไซต์ kaggle.com โดยจะมีตัวแปรที่ใช้ทั้งหมด 2 ตัวแปรคือ ข้อความและหมายเลขคำตอบของข้อมูล ทั้งหมด 5,171 ข้อความ ทำการจัดเตรียมข้อมูลโดยจะนำข้อมูลที่ผ่านการจัดเตรียมข้อมูลไปกำจัดค่านอกเกณฑ์และแบ่งข้อมูลอีกชุดที่ไม่ได้ทำการกำจัดค่านอกเกณฑ์ แล้วนำข้อมูลที่กำจัดค่านอกเกณฑ์และไม่กำจัดค่านอกเกณฑ์มาสร้างแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก จากนั้นใช้วิธีการเรียนรู้แบบรวมกลุ่มในการเปรียบเทียบประสิทธิภาพการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์ โดยสามารถแสดงการสรุปผลการวิจัยและข้อเสนอแนะได้ดังนี้

5.1 สรุปผลการวิจัย

5.1.1 จากการกำจัดค่านอกเกณฑ์ทั้ง 3 วิธี คือ วิธี DBSCAN วิธีป่าไม้โดดเดี่ยว และวิธี IF-LOF โดยจะใช้ข้อมูลที่ผ่านการจัดเตรียมข้อมูลมาแล้วมาใช้เป็นข้อมูลเข้า ซึ่งวิธี DBSCAN มีการตรวจจับว่าเป็นค่านอกเกณฑ์ 20 คำ เป็นค่าปกติ 45,067 คำ วิธีป่าไม้โดดเดี่ยว มีการตรวจจับว่าเป็นค่านอกเกณฑ์ 902 คำ เป็นค่าปกติ 44,185 คำ และวิธี IF-LOF มีการตรวจจับว่าเป็นค่านอกเกณฑ์ 46 คำ เป็นค่าปกติ 856 คำ แล้วนำมาเข้าแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก ซึ่งจะพบว่าวิธี DBSCAN ในแต่ละแบบจำลองมีค่าความแม่น ค่าความเที่ยง ค่าเรียกคืน และค่าประสิทธิภาพโดยรวมสูงที่สุด

5.1.2 การสร้างแบบจำลองการเรียนรู้ของเครื่อง ทั้ง 3 วิธี คือ วิธีนาอีฟเบส์ วิธีซัพพอร์ตเวกเตอร์แมชชีน และวิธีเพื่อนบ้านใกล้สุด k ตัว แล้วใช้วิธีการเรียนรู้แบบรวมกลุ่มกับการเรียนรู้ของเครื่อง โดยจะใช้ข้อมูลที่มีการกำจัดค่านอกเกณฑ์และไม่กำจัดค่านอกเกณฑ์มาใช้เป็นข้อมูลเข้า แล้วเปรียบเทียบค่าความแม่น ค่าความเที่ยง ค่าเรียกคืน และค่าประสิทธิภาพโดยรวมในวิธีการเรียนรู้แบบรวมกลุ่มในแต่ละข้อมูลที่มีการกำจัดค่านอกเกณฑ์และไม่กำจัดค่านอกเกณฑ์

การเรียนรู้ของเครื่องด้วยวิธีนาอีฟเบส์ใช้การกำจัดค่านอกเกณฑ์ให้ประสิทธิภาพดีที่ดีกว่าข้อมูลที่ไม่กำจัดค่านอกเกณฑ์ โดยวิธี DBSCAN ให้ประสิทธิภาพการทำนายดีที่สุดเนื่องจากมีค่าความแม่น ค่าเรียกคืน และค่าประสิทธิภาพโดยรวมสูงที่สุด 97.55% 94.00% และ 95.70% ตามลำดับรองลงมาเป็นวิธี IF-LOF และวิธีป่าไม้โดดเดี่ยว

การเรียนรู้ของเครื่องด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนใช้การกำจัดค่านอกเกณฑ์ให้ประสิทธิภาพดีที่ดีกว่าข้อมูลที่ไม่กำจัดค่านอกเกณฑ์ โดยวิธี DBSCAN ให้ประสิทธิภาพการทำนายดีที่สุดเนื่องจากมีค่าความแม่น ค่าเรียกคืน และค่าประสิทธิภาพโดยรวมสูงที่สุด ฟังก์ชันเคอร์เนลที่นำมาใช้คือฟังก์ชันเกาส์เซียนเรเดียลเบสิสเนื่องจากมีค่าความแม่น ค่าเรียกคืน และค่าประสิทธิภาพโดยรวมสูงที่สุด 97.29% 96.89% และ 95.40% ตามลำดับรองลงมาเป็นวิธี IF-LOF และวิธีป่าไม้โดดเดี่ยว

การเรียนรู้ของเครื่องด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวใช้การกำจัดค่านอกเกณฑ์ให้ประสิทธิภาพดีที่ดีกว่าข้อมูลที่ไม่กำจัดค่านอกเกณฑ์ โดยวิธี DBSCAN ให้ประสิทธิภาพการทำนายดีที่สุดโดยกำหนดค่า k เท่ากับ 1 เนื่องจากมีค่าความแม่น ค่าเรียกคืน และค่าประสิทธิภาพโดยรวมสูงที่สุด 86.02% 96.44% และ 66.21% ตามลำดับรองลงมาเป็นวิธี IF-LOF และวิธีป่าไม้โดดเดี่ยว

แบบจำลองการเรียนรู้ของเครื่องใช้การกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN ให้ประสิทธิภาพดีที่สุดในทุกแบบจำลอง การเรียนรู้ของเครื่องด้วยวิธีนาอีฟเบส์ให้ค่าประสิทธิภาพในการทำนายดีที่สุด

วิธีการเรียนรู้แบบรวมกลุ่มของการเรียนรู้ของเครื่องด้วยวิธีการโหวตเสียงข้างมาก ใช้การกำจัดค่านอกเกณฑ์ให้ประสิทธิภาพดีที่ดีกว่าข้อมูลที่ไม่กำจัดค่านอกเกณฑ์ โดยวิธี DBSCAN ให้ประสิทธิภาพการทำนายดีที่สุดเนื่องจากมีค่าความแม่น ค่าเรียกคืน และค่าประสิทธิภาพโดยรวมสูงที่สุด 98.13% 98.22% และ 96.82% ตามลำดับลองลงมาเป็นวิธี IF-LOF และวิธีป่าไม้โดดเดี่ยว

5.1.3 การสร้างแบบจำลองการเรียนรู้เชิงลึกทั้ง 3 วิธี คือวิธีโครงข่ายประสาทเทียม วิธีโครงข่ายประสาทแบบวนซ้ำ และวิธีหน่วยความจำระยะสั้นยาวแล้วใช้วิธีการเรียนรู้แบบรวมกลุ่มกับการเรียนรู้เชิงลึก โดยจะใช้ข้อมูลที่มีการกำจัดค่านอกเกณฑ์และไม่กำจัดค่านอกเกณฑ์มาใช้เป็นข้อมูลเข้า แล้วเปรียบเทียบค่าความแม่น ค่าความเที่ยง ค่าเรียกคืน และค่าประสิทธิภาพโดยรวมในวิธีการเรียนรู้แบบรวมกลุ่มในแต่ละข้อมูลที่มีการกำจัดค่านอกเกณฑ์และไม่กำจัดค่านอกเกณฑ์

การเรียนรู้เชิงลึกด้วยวิธีโครงข่ายประสาทเทียมใช้การกำจัดค่านอกเกณฑ์ให้ประสิทธิภาพดีที่ดีกว่าข้อมูลที่ไม่กำจัดค่านอกเกณฑ์ โดยวิธี DBSCAN ให้ประสิทธิภาพการทำนายดีที่สุดเนื่องจากมีค่าความแม่น ค่าเรียกคืน และค่าประสิทธิภาพโดยรวมสูงที่สุด 98.26% 99.11% และ 97.06% ตามลำดับรองลงมาเป็นวิธี IF-LOF และวิธีป่าไม้โดดเดี่ยว

การเรียนรู้เชิงลึกด้วยวิธีโครงข่ายประสาทแบบวนซ้ำใช้การกำจัดค่านอกเกณฑ์ให้ประสิทธิภาพดีที่ดีกว่าข้อมูลที่ไม่กำจัดค่านอกเกณฑ์ โดยวิธี DBSCAN ให้ประสิทธิภาพการทำนายดีที่สุดเนื่องจากมีค่าความแม่น ค่าเรียกคืน และค่าประสิทธิภาพโดยรวมสูงที่สุด 95.04% 87.56% และ 91.09% ตามลำดับรองลงมาเป็นวิธี IF-LOF และวิธีป่าไม้โดดเดี่ยว

การเรียนรู้เชิงลึกด้วยวิธีหน่วยความจำระยะสั้นยาวใช้การกำจัดค่านอกเกณฑ์ให้ประสิทธิภาพดีที่ดีกว่าข้อมูลที่ไม่กำจัดค่านอกเกณฑ์ โดยวิธี DBSCAN ให้ประสิทธิภาพการทำนายดีที่สุดเนื่องจากมีค่าความแม่น ค่าเรียกคืน และค่าประสิทธิภาพโดยรวมสูงที่สุด 97.81% 98.15% 94.22% และ 96.15% ตามลำดับรองลงมาเป็นวิธี IF-LOF และวิธีป่าไม้โดดเดี่ยว

แบบจำลองวิธีการเรียนรู้เชิงลึกใช้การกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN ให้ผลลัพธ์ดีที่สุดในทุกแบบจำลอง การเรียนรู้เชิงลึกด้วยวิธีโครงข่ายประสาทเทียมให้ค่าประสิทธิภาพในการทำนายดีที่สุด

วิธีการเรียนรู้แบบรวมกลุ่มของการเรียนรู้เชิงลึกด้วยวิธีการโหวตเสียงข้างมาก ไม่ใช้การกำจัดค่านอกเกณฑ์ให้ประสิทธิภาพดีที่ดีกว่าข้อมูลที่มีการกำจัดค่านอกเกณฑ์ให้ประสิทธิภาพการทำนายดีที่สุดเนื่องจากมีค่าความแม่น ค่าความเที่ยง ค่าเรียกคืน และค่าประสิทธิภาพโดยรวมสูงที่สุด 98.52% 97.55% 97.33 และ 97.44% เนื่องจากอาจทำให้การวิเคราะห์ผิดพลาดเพราะยังมีค่านอกเกณฑ์อยู่ จึงใช้ประสิทธิภาพการทำนายของวิธีการเรียนรู้แบบรวมกลุ่มที่ทำการกำจัดค่านอกเกณฑ์มาเปรียบเทียบประสิทธิภาพ

5.1.4 เปรียบเทียบแบบจำลองการเรียนรู้แบบรวมกลุ่มกับวิธีการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์โดยใช้การกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN

ผลลัพธ์ประสิทธิภาพการทำนายของแบบจำลองการเรียนรู้แบบรวมกลุ่มของการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก โดยใช้ข้อมูลที่ผ่านการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN จะสรุปได้ว่าการเรียนรู้เชิงลึกให้ประสิทธิภาพการทำนายได้ดีกว่าแบบจำลองการเรียนรู้ของเครื่อง เนื่องจากค่าความแม่น ค่าความเที่ยง และค่าประสิทธิภาพโดยรวมมากกว่า

5.2 ข้อจำกัดและข้อเสนอแนะ

5.2.1 ข้อจำกัด

1) เนื่องจากเวลาในการทำวิจัยที่จำกัด จึงไม่สามารถนำการกำจัดค่านอกเกณฑ์ การเรียนรู้ของเครื่อง และการเรียนรู้เชิงลึกวิธีอื่นมาทำการทดสอบได้ซึ่งอาจจะมีวิธีที่ให้ค่าประสิทธิภาพการทำนายที่ดีกว่า

2) การสร้างแบบจำลองบางตัวจะต้องใช้เวลาในการทำนายที่นาน อุปกรณ์ที่ใช้ในการวิเคราะห์ต้องมีความพร้อมและมีประสิทธิภาพในการทดสอบที่สูงกว่านี้

5.2.2 ข้อเสนอแนะ

1) เนื่องจากข้อมูลที่ใช้เป็นข้อมูลจากเว็บไซต์ kaggle มีเพียงชุดข้อมูลเดียว จึงควรหาชุดข้อมูลจดหมายอิเล็กทรอนิกส์มาเพิ่มจากแหล่งอื่น เพื่อให้ชุดข้อมูลมีความหลากหลายมากขึ้น ไม่ได้มาจากแหล่งเดียว จะช่วยเพิ่มให้แบบจำลองมีการเรียนรู้จากชุดข้อมูลที่มีความหลากหลาย

2) ควรมีการทดสอบแบบจำลองที่มากกว่านี้เพื่อพิจารณาว่าแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกวิธีไหนมีประสิทธิภาพดีกว่ากันในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์ เช่น วิธีต้นไม้ตัดสินใจ (Decision Tree), วิธี AdaBoost และ วิธีหน่วยเวียนกลับแบบมีประตู (Gated Recurrent Unit) เป็นต้น

3) ในวิธีการเรียนรู้เชิงลึกอาจจะต้องมีการปรับค่าพารามิเตอร์ที่มากกว่านี้ เพื่อหาค่าที่ดีที่สุด แต่จะต้องใช้เวลาในการทำนายที่นานขึ้น

4) สามารถนำไปต่อยอดในการตรวจจับเว็บไซต์ปลอม เนื่องจากปัจจุบันการส่งจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์เริ่มมีน้อยส่วนใหญ่นิยมส่งเป็นเว็บไซต์ปลอมแทน

เอกสารอ้างอิง

กฤษฎา ออยพันธุดี. 2556. ความผิดเกี่ยวกับการส่งข้อมูลหรือจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์: ศึกษากฎหมายของประเทศไทยและกฎหมายต่างประเทศ. นิติศาสตรมหาบัณฑิต สาขาวิชานิติศาสตร. มหาวิทยาลัยธุรกิจบัณฑิตย์.

กาญจนา ทองบุญนาค. 2561. การพัฒนาแบบจำลองเพื่อพยากรณ์ปริมาณ PM10 ในพื้นที่จังหวัดเชียงใหม่ โดยใช้โครงข่ายเพอร์เซ็ปตรอนหลายชั้น. มหาวิทยาลัยราชภัฏเชียงใหม่.

ณัฏฐณิชา ยงยิ่ง. 2562. การประยุกต์ใช้เทคโนโลยีการเรียนรู้เชิงลึกในการจำแนกข้อมูลถนนจากภาพถ่าย Drone เพื่อการสำรวจถนนในเขตชนบท. วิทยาศาสตรบัณฑิต สาขาวิชาภูมิศาสตร์. มหาวิทยาลัยนเรศวร.

เดช ธรรมศิริ และพยุง มีสัจ. 2554. การจำแนกข้อมูลด้วยวิธีแบบร่วมกันตัดสินใจจากพื้นฐานของเทคนิคต้นไม้ตัดสินใจเทคนิคโครงข่ายประสาทเทียมและเทคนิคซัพพอร์ตเวกเตอร์แมชชีนร่วมกับการเลือกตัวแทนที่เหมาะสมด้วยขั้นตอนวิธีเชิงพันธุกรรม. วารสารวิชาการพระจอมเกล้าพระนครเหนือ. 2: 293-303

ธนาภัทร ภัทรวินิจ. 2563. การทำนายความผิดพลาดระยะต้นของเครื่องวิเคราะห์อินทรีย์คาร์บอนโดยการเรียนรู้เชิงลึก. วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์. จุฬาลงกรณ์มหาวิทยาลัย.

เธียรศักดิ์ พลาดิศัยเลิศ และธนิศา นุ่มนนท์. 2561. การเปรียบเทียบผลการทำนายราคาบิทคอยน์ด้วยการเรียนรู้ของเครื่องแบบต่างๆ. วารสารเทคโนโลยีสารสนเทศลาดกระบัง. 1: 1-9

บุษบงก์ คชินทรโรจน์, เดือนเพ็ญ ธีรวรรณวิวัฒน์ และพาชิตชนัต ศิริพานิช. 2564. การสร้างระบบคัดกรองข้อความการเกลียดกลัวคนต่างชาติบนทวิตเตอร์ในช่วงการแพร่ระบาดของโรคติดเชื้อไวรัสโคโรนา 2019. วารสารไทยการวิจัยดำเนินงาน. 1: 31-44.

ปรเมษฐ์ ธันวานนท์, ชัยกร ยิ่งเสรี, วรพล พงษ์เพ็ชร และธนภัทร ฆังคะจิตร. 2560. การประยุกต์ใช้โมเดลการเรียนรู้แบบรวมกลุ่มเพื่อพยากรณ์แนวโน้มของราคาหลักทรัพย์ในตลาดหลักทรัพย์แห่งประเทศไทย. วารสารวิทยาการและเทคโนโลยีสารสนเทศ. 1: 12-21. doi 10.14456/jist.2017.2.

พัชนา สุวรรณแสน. 2562. การจัดการข้อมูลสูญหาย: วิธีเคเนียร์เรสเนเบอร์. วารสารวิจัยวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครราชสีมา. 1: 1-9.

ภัครพล อาจอาษา. 2564. การวิเคราะห์คุณภาพน้ำด้วยเทคนิคการจัดกลุ่มข้อมูล. วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ. มหาวิทยาลัยมหาสารคาม.

เอกสารอ้างอิง (ต่อ)

วิศรุต แก้วมหา และวริศ ปัญญาฉัตรพร. 2563. การคาดการณ์ผลตอบแทนในอนาคตของตราสารทุนหุ้นสามัญโดยการใช้ระบบคอมพิวเตอร์เรียนรู้ได้ด้วยตนเอง. วารสารนวัตกรรมธุรกิจ การจัดการ และสังคมศาสตร์. 3: 108-123.

วิศาล พัฒน์ชู. 2549. ขยะไปรษณีย์อิเล็กทรอนิกส์ (Spam Mail). Innovation and Technology. 88-93.

วีระพันธ์ พานิชย์. 2564. การประยุกต์ใช้ Machine Learning ทำนายผลการเรียนวิชา Web Database. วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์และโทรคมนาคม. มหาวิทยาลัยธุรกิจบัณฑิตย์.

สุจิรา ไชยกุสินธุ์. 2560. การประยุกต์ใช้วิธีการ KNN สำหรับกลยุทธ์การซื้อขายหุ้น. มหาวิทยาลัยราชมงคลพระนคร.

สุทธิพงศ์ ผ่องแผ้ว. 2561. การศึกษากระบวนการแบ่งกลุ่มเมล็ดพันธุ์ด้วยข้อมูลโครงสร้างของเมล็ด. วิทยาศาสตรบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ. มหาวิทยาลัยศรีนครินทรวิโรฒ.

อนัตต์ชัย ชุติภาสเจริญ และจรัญ แสนราช. 2561. การเปรียบเทียบประสิทธิภาพของอัลกอริทึมและการคัดเลือกคุณลักษณะที่เหมาะสมเพื่อการพยากรณ์โอกาสความสำเร็จในการโอนเงินข้ามประเทศของบุคคลทั่วไป. วารสารวิจัย มข. (ฉบับบัณฑิตศึกษา) สาขามนุษยศาสตร์และสังคมศาสตร์. 3: 105-113.

อัศวิน สุรวัชโยธิน และวรภัทร ไพรีเกรง. 2564. การสร้างตัวแบบการทำนายในการเลือกศึกษาต่อในระดับอุดมศึกษาโดยการใช้เทคนิคแบบบูรณาการในการแก้ปัญหาการจำแนกข้อมูลไม่สมดุลของกลุ่มผู้เรียน. วารสารวิทยาการและเทคโนโลยีสารสนเทศ. 1: 65-79

อาณัตชัย เตชะวิเศษชัย. 2564. Anomaly Detection with Isolation Forest: แยกข้อมูลผิดปกติง่ายๆ ด้วย Isolation Forest. [Online]. เข้าถึงได้จาก https://bigdataexperience.org.

เอกพันธ์ บุญเสริม. 2563. การประยุกต์ใช้การเรียนรู้ของเครื่องในการทำนายความรุนแรงของผู้บาดเจ็บจากอุบัติเหตุทางถนนในช่วงเทศกาลปีใหม่จากข้อมูลเปิดภาครัฐของประเทศไทย. วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ. มหาวิทยาลัยศรีนครินทรวิโรฒ.

Awad, W. and Elseuofi, S. 2011. Machine learning method for spam e-mail classification. International Journal of Computer Science & Information Technology (IJCSIT). 1: 173-184. doi 10.5121/ijcsit.2011.3112.

เอกสารอ้างอิง (ต่อ)

Dietrich, D., Heller, B. and Yang, B. 2015. Data Science & Big Data Analytics Discovering, Analyzing, Visualizing and Presenting Data. Indiana. John Wiley & Sons,Inc.

Farzad, A. and Gulliver, T. 2020. Unsupervised log message anomaly detection. International ICT Express. 6: 229-237. doi 10.1016/j.icte.2020.06.003.

Hossain, F., Uddin, M. and Halder, R. 2021. Analysis of optimized machine learning and deep learning techniques for spam detection. In Electronics and Mechatronics Conference (IEMTRONICS). Toronto.

Julis, M. and Alagesan, S. 2020. Spam detection in sms using machine learning through text mining. International Journal of Scientific & Technology Research. 2: 498-503.

Louis, H., Stuti, T., Louis, T., Mengyang, C. and Padmini, S. 2022. Text preprocessing for text mining in organizational research: review and recommendations. Organizational Research Methods. 25: 114-146. doi 10.1177/1094428120971683.

Mountrakis, G., Im, J. and Ogole, C. 2011. Support vector machines in remote sensing: A review. ISPRS Journal of Photogrammetry and Remote Sensing. 66: 247-259.

Poomka, P., Pongsena, W., Kerdprasop, W. and Kerdprasop, K. 2019. SMS spam detection based on long short-term memory and gated recurrent unit. International Journal of Future Computer and Communication. 1: 11-15. doi 10.18178/ijfcc.2019.8.1.532.

Vijayarani, S., Ilamathi J., Nithya. 2015. Preprocessing techniques for text mining - an overview. International Journal of Computer Science & Communication Networks. 5: 7-16.

Wang, S., Mathew, D., Chen, Y., Xi, L., Ma, L. and Lee, J. 2009. Empirical analysis of support vector machine ensemble classifiers. Expert Systems with Applications. 36: 6466-6476.

Zhangyu, C., Chengming, Z. and Jianwei, D. 2019. Outlier detection using isolation forest and local outlier factor. 161-168. Proceedings of International Conference on Research in Adaptive and Convergent Systems. Chongqing. doi 10.1145/3338840.3355641.

ภาคผนวก

ภาคผนวก ก

ภาคผนวก ก.1 ตัวอย่างข้อมูลที่ไม่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์

	CleanedText	label
0	enron methanol meter follow note gave monday p...	0
1	hpl nom january see attached file hplnol xls h...	0
2	neon retreat ho ho ho around wonderful time ye...	0
3	photoshop windows office cheap main trending a...	1
4	indian springs deal book teco pvr revenue unde...	0
...	...	...
5166	put ft transport volumes decreased contract th...	0
5167	following noms hpl take extra mmcf weekend try...	0
5168	calpine daily gas nomination julie mention ear...	0
5169	industrial worksheets august activity attached...	0
5170	important online banking alert dear valued cit...	1

ภาคผนวก ก.2 ตัวอย่างข้อมูลที่กำจัดคำที่เป็นค่านอกเกณฑ์ของวิธี DBSCAN

	DB	label
0	enron methanol follow note gave monday prelimi...	0
1	nom january see attached file hplnol hplnol	0
2	neon retreat around wonderful time year neon l...	0
3	photoshop windows office cheap main trending a...	1
4	indian springs book teco pvr revenue understan...	0
...	...	...
5166	put ft transport volumes decreased contract th...	0
5167	following noms take extra mmcf weekend try nex...	0
5168	calpine daily gas nomination julie mention ear...	0
5169	industrial worksheets august activity attached...	0
5170	important online banking alert dear valued cit...	1

ภาคผนวก ก.3 ตัวอย่างข้อมูลที่กำจัดคำที่เป็นค่านอกเกณฑ์ของวิธีป่าไม้โดดเดี่ยว

	IF	label
0	gave preliminary override presently zero refle...	0
1	hplnol hplnol	0
2	neon retreat wonderful neon leaders retreat ex...	0
3	main trending abasements darer prudently fortu...	1
4	indian springs pvr revenue understanding sends...	0
...	...	...
5166	decreased royal edmondson	0
5167	extra mmcf weekend stay mmcf fcv mcmills ftwor...	0
5168	mention earlier afternoon experiencing difficu...	0
5169	industrial worksheets worksheets different wor...	0
5170	banking alert dear valued citizensr concerns s...	1

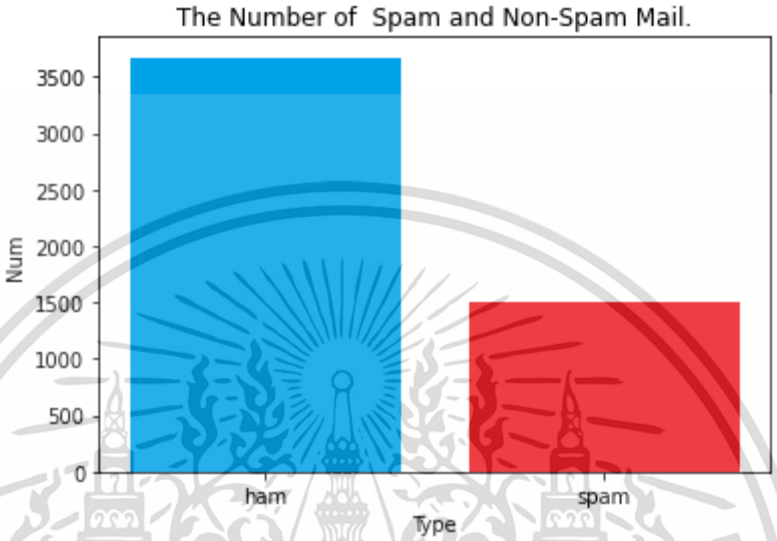
ภาคผนวก ก.4 ตัวอย่างข้อมูลที่กำจัดคำที่เป็นค่านอกเกณฑ์ของวิธี IF-LOF

	lof	label
0	methanol follow note gave monday preliminary f...	0
1	nom january see attached file hplnol hplnol	0
2	neon retreat around wonderful time year neon l...	0
3	photoshop windows office cheap main trending a...	1
4	indian springs book teco pvr revenue understan...	0
...	...	...
5166	put ft transport volumes decreased contract th...	0
5167	following noms take extra mmcf weekend try nex...	0
5168	calpine daily gas nomination julie mention ear...	0
5169	industrial worksheets august activity attached...	0
5170	important banking alert dear valued citizensr ...	1

ภาคผนวก ข

ภาคผนวก ข.1 การวิเคราะห์จำนวนข้อความจดหมายอิเล็กทรอนิกส์

จากการนำข้อมูลจดหมายอิเล็กทรอนิกส์จากเว็บไซต์ kaggle.com ทั้งหมด 5,171 ข้อความ โดยจำนวนจดหมายอิเล็กทรอนิกส์ปกติและจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์ แสดงดังรูป 4.1

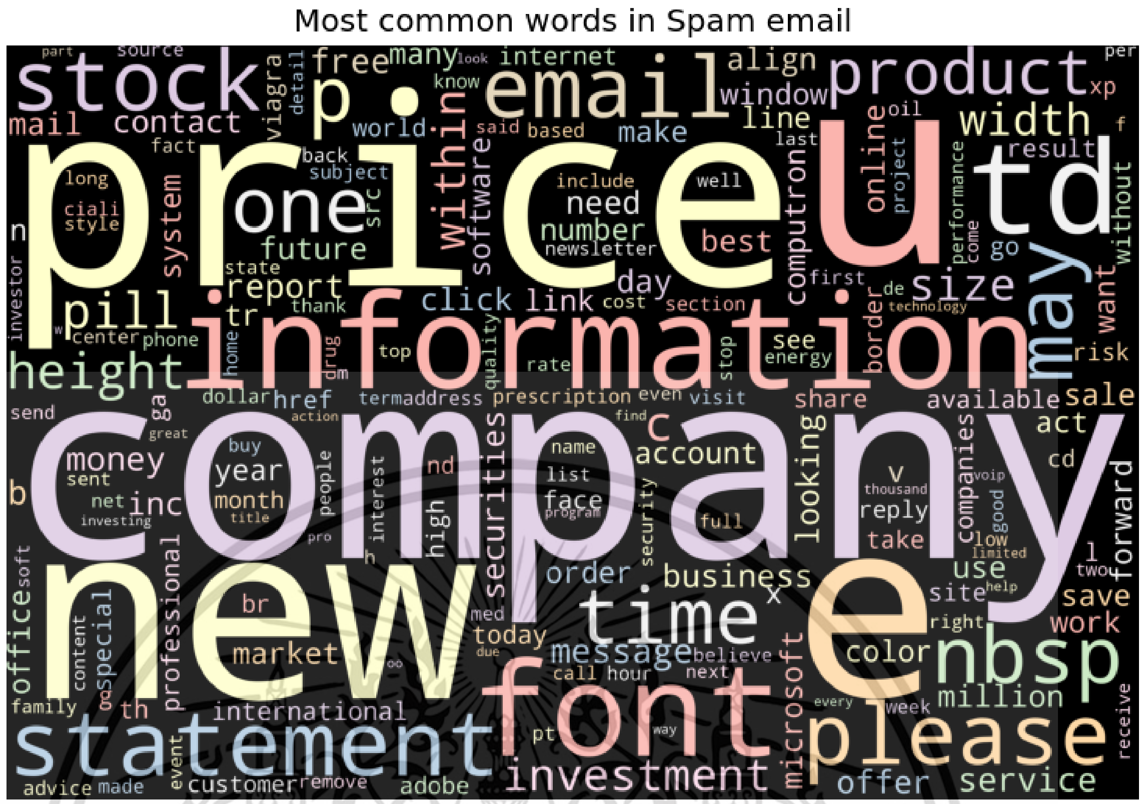


รูปที่ ข.1 จำนวนจดหมายอิเล็กทรอนิกส์ปกติและจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์

จากรูปที่ ข.1 แสดงจำนวนจดหมายอิเล็กทรอนิกส์ประกอบด้วยจดหมายอิเล็กทรอนิกส์ปกติ (ham) มี 3,672 ข้อความ คิดเป็น 71.01% และจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์ (spam) มี 1,499 ข้อความ คิดเป็น 28.99%

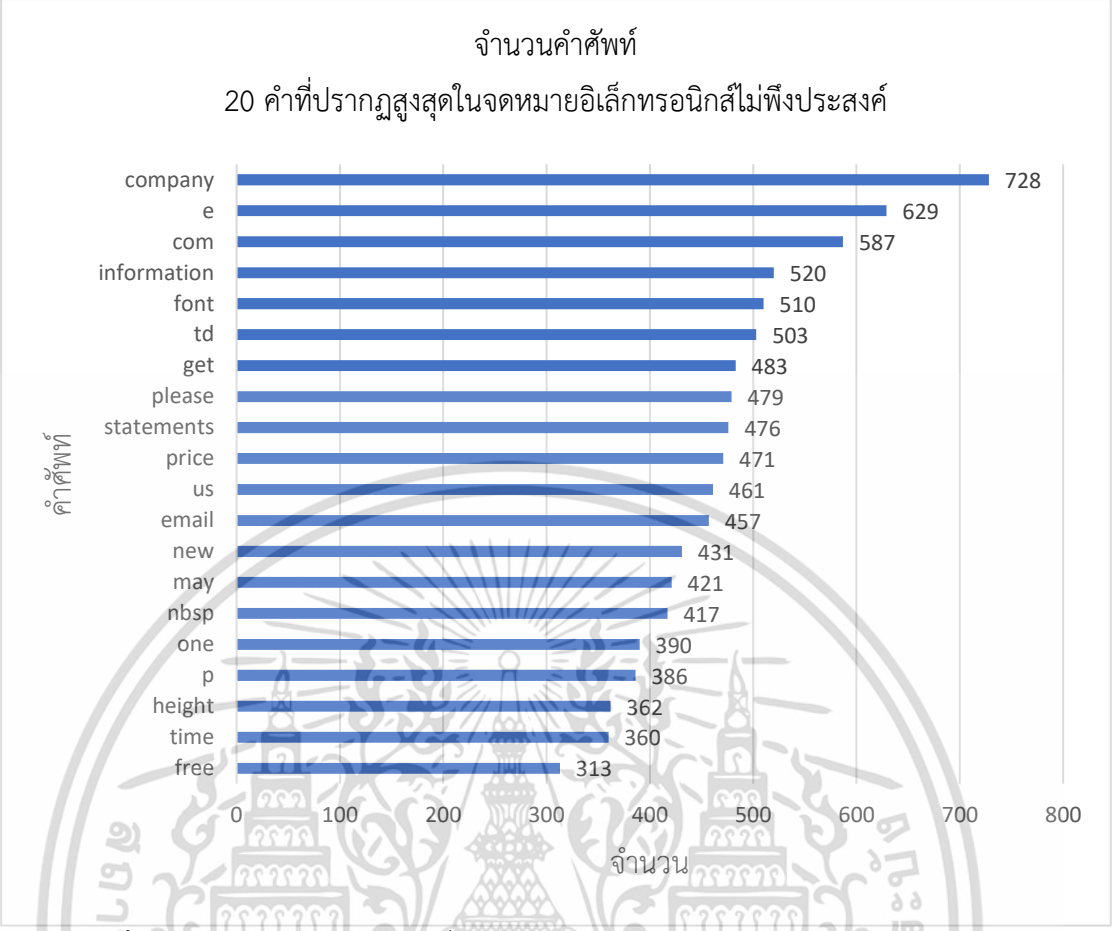
ภาคผนวก ข.2 ผลการวิเคราะห์ความถี่ของคำในจดหมายอิเล็กทรอนิกส์

ผู้วิจัยได้ทำการสร้างกราฟแจกแจงความถี่ของคำในจดหมายอิเล็กทรอนิกส์และสร้างกราฟกลุ่มคำที่แสดงขนาดตามจำนวนความถี่ของคำ (WordCloud) และกราฟแท่ง (Bar Chart) เพื่อให้อยู่ในรูปแบบที่ดูง่าย



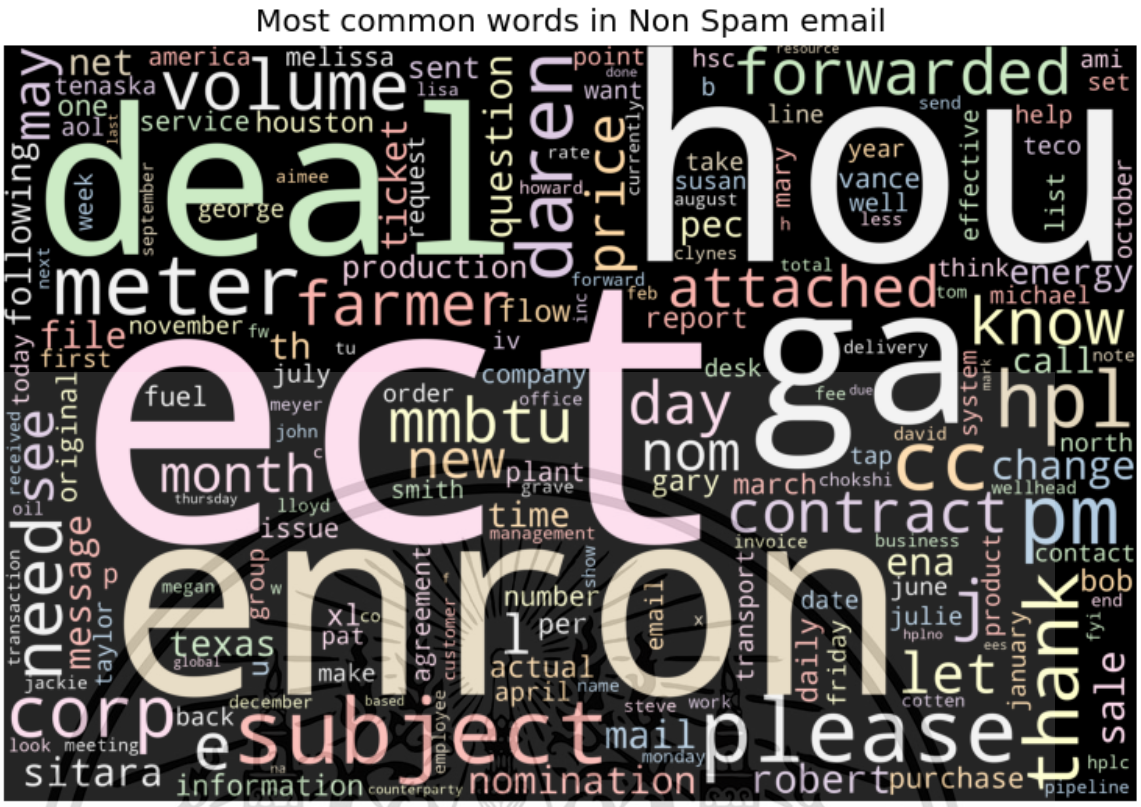
รูป ข.2 รูปความถี่ของคำ (WordCloud) ของจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์

จากรูปที่ ข.2 แสดงผลลัพธ์ของคำในจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์ในรูปแบบของการแสดงความถี่ของคำ โดยคำที่มีขนาดใหญ่จะเป็นคำที่ปรากฏบ่อยในจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์ ประกอบด้วยคำว่า “company” “price” “u” และ “new” “information” เป็นต้น



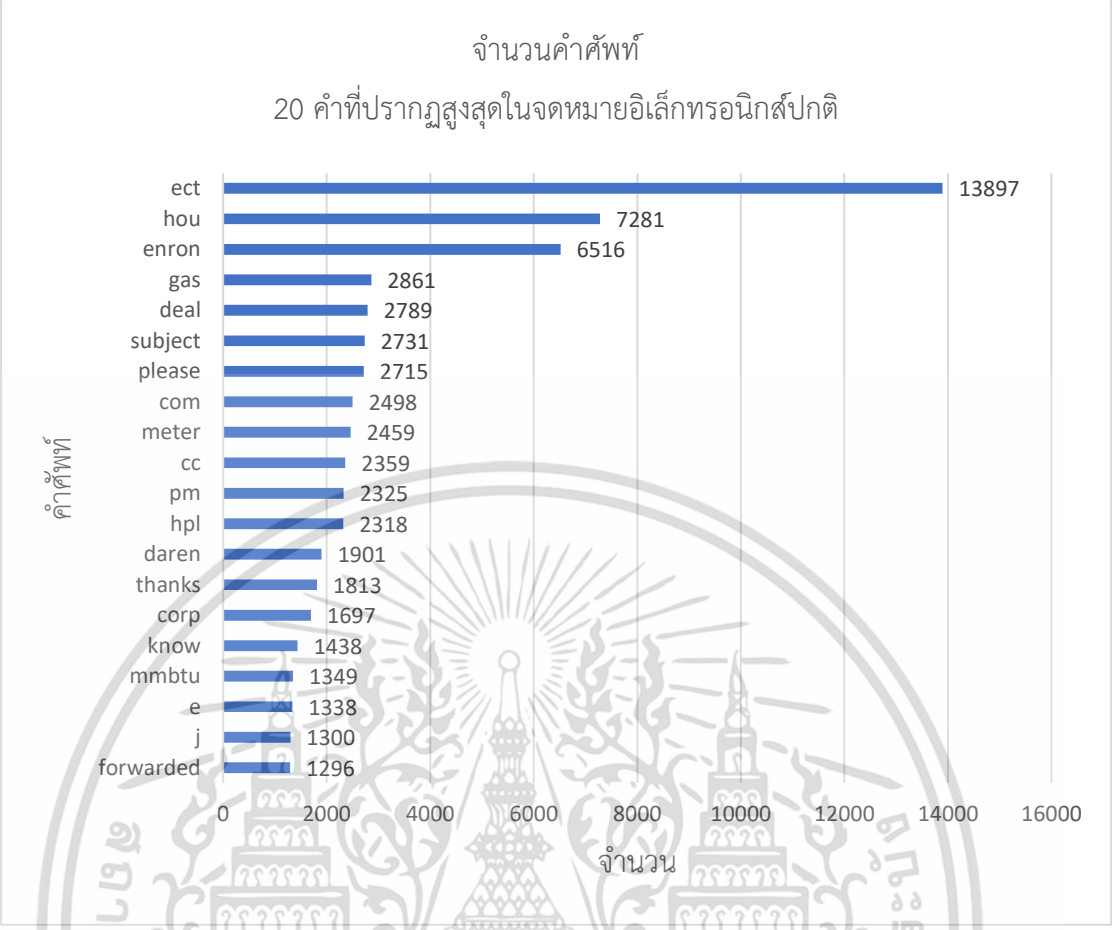
รูปที่ ข.3 จำนวนคำศัพท์ 20 คำที่ปรากฏสูงสุดในจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์

รูปที่ ข.3 แสดง 20 อันดับแรกของคำที่ปรากฏสูงสุดในจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์ ซึ่งประกอบด้วย “company” “e” “com” “information” “font” และ “td” เป็นต้น จากจำนวนคำทั้งหมด 36,608 คำ



รูปที่ ข.4 รูปความถี่ของคำ (WordCloud) ของจดหมายอิเล็กทรอนิกส์ปกติ

จากรูปที่ ข.4 แสดงผลลัพธ์ของคำในจดหมายอิเล็กทรอนิกส์ปกติในรูปแบบของการแสดงความถี่ของคำ โดยคำที่มีขนาดใหญ่จะเป็นคำที่ปรากฏบ่อยในจดหมายอิเล็กทรอนิกส์ปกติ ประกอบด้วยคำว่า “ect” “hou” “deal” “enron” “ga” และ “please” เป็นต้น



รูปที่ ข.5 จำนวนคำศัพท์ 20 คำที่ปรากฏสูงสุดในจดหมายอิเล็กทรอนิกส์ปกติ

รูปที่ 4.5 แสดง 20 อันดับแรกของคำที่ปรากฏสูงสุดในจดหมายอิเล็กทรอนิกส์ปกติ ซึ่งประกอบด้วย “ect” “hou” “enron” “gas” “deal” และ “subject” เป็นต้น จากจำนวนคำทั้งหมด 16,028 คำ

ภาคผนวก ค

ภาคผนวก ค ชุดคำสั่งไพทอน (Python) ที่ใช้ในการเก็บรวบรวมข้อมูล การตัดคำภาษาอังกฤษการกำจัดค่านอกเกณฑ์และการสร้างแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกในการตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์

ภาคผนวก ค.1 ชุดคำสั่งที่ใช้ในการวิจัย (Library)

Numpy	NumPy เป็นชุดคำสั่งที่ใช้ในการคำนวณทางคณิตศาสตร์ด้วยภาษา Python ซึ่งสามารถคำนวณ และดำเนินการทางตรรกะใน Array หรือ Matrix ได้อย่างรวดเร็ว
Pandas	Pandas คือหนึ่งในชุดคำสั่งสำคัญของภาษา Python มีความสามารถในการจัดการ และวิเคราะห์ข้อมูลได้อย่างมีประสิทธิภาพตั้งแต่ข้อมูลขนาดเล็กไปจนถึงข้อมูลขนาดใหญ่สามารถใช้การเขียนโค้ด เพื่อปรับแต่ง หรือเชื่อมต่อกับโปรแกรมอื่นๆเพื่อดูชุดข้อมูล
nltk	Natural Language Toolkit หรือเรียกว่า nltk เป็นการประมวลภาษาธรรมชาติ เป็นส่วนหนึ่งของปัญญาประดิษฐ์และภาษาศาสตร์ เพื่อให้คอมพิวเตอร์สามารถตีความและเข้าใจภาษามนุษย์ได้
String	เป็นประเภทของข้อมูลที่เป็นข้อความโดยประกอบไปด้วยหลายตัวอักษรข้อมูลซึ่งใช้สำหรับเก็บข้อมูลหนึ่งตัวอักษร ดังนั้นในการที่จะเก็บหลายตัวอักษร เราจะต้องใช้ความสามารถของอาเรย์เข้ามาช่วยเพื่อทำงานร่วมกัน

re	เครื่องมือที่ใช้ดึงข้อมูลที่ต้องการจาก string ผ่านการกำหนด pattern ของข้อมูลที่ต้องการดึงโดยที่ไม่ต้องระบุข้อมูลที่ต้องการ ทำให้เราสามารถดึงข้อมูลที่ไม่ตายตัว แต่มี pattern ชัดเจนได้ เช่น email ที่มีการระบุ @, เบอร์โทรศัพท์ที่มีจำนวนตัวเลขคงที่ และอื่นๆ
matplotlib	Matplotlib เป็นชุดคำสั่งของภาษา Python เพื่อใช้ในการสร้างหรือแสดงผล Data visualization ช่วยในการสร้างแผนภูมิและกราฟต่างๆเพื่อช่วยในการวิเคราะห์ที่ทำให้ดูง่ายขึ้น
seaborn	Seaborn เป็นชุดคำสั่งสำหรับสร้างกราฟิกทางสถิติใน Python สร้างขึ้นจากคำสั่ง matplotlib
sklearn	Scikit-learn หรือ sklearn เป็นชุดคำสั่งของภาษา Python ใช้สำหรับการเรียนรู้ของเครื่องและสร้างตัวแบบทางสถิติ การจำแนกประเภท เช่น Regression, Classification และ Clustering
Gensim	Gensim เป็นชุดคำสั่งของภาษา Python ช่วยในการจัดการแปลงคำเป็นตัวเลขด้วย Word2Vec

ภาคผนวก ค.2 ชุดคำสั่งไพทอนในการนำข้อมูลเข้า

```
# ติดตั้ง package
import pandas as pd
# นำข้อมูลเข้ามาเก็บไว้ในตัวแปร
mail_data = pd.read_csv('ตำแหน่งที่เก็บไฟล์')
```

ภาคผนวก ค.3 ชุดคำสั่งไพทอนในการลบสดมภ์ No. และ label

```
# ลบตัวแปร No. และ label
mail_data.drop('Unnamed: 0', axis=1, inplace = True)
mail_data.drop('label', axis=1, inplace = True)
```

ภาคผนวก ค.4 ชุดคำสั่งไพทอนในการเปลี่ยนชื่อสดมภ์

```
# ในการเปลี่ยนชื่อตัวแปร label_num เป็น label
mail_data.columns = ['text','label']
```

ภาคผนวก ค.5 ชุดคำสั่งไพทอนในการสร้างสดมภ์ใหม่

```
#สร้างสดมภ์ใหม่ชื่อ CleanedText
Cleaned = mail_data['text']
CleanedText = pd.DataFrame(Cleaned)
mail_data['CleanedText']=CleanedText
```

ภาคผนวก ค.6 ชุดคำสั่งไพทอนในการเปลี่ยนตัวอักษรพิมพ์ใหญ่เป็นพิมพ์เล็ก

```
#เปลี่ยนตัวอักษร
mail_data['CleanedText'] = mail_data['CleanedText'].str.lower()
```

ภาคผนวก ค.7 ชุดคำสั่งไพทอนในการลบคำหยุด

```
#ลบคำหยุด (Stopwords)
STOPWORDS = set(stopwords.words('english'))
def remove_stopwords(maill):
    return " ".join([word for word in maill.split() if word not in STOPWORDS])
mail_data['CleanedText'] = mail_data['CleanedText'].apply(remove_stopwords)
```

ภาคผนวก ค.8 ชุดคำสั่งไพทอนในการลบลิงก์ที่นำไปสู่เว็บไซต์

```
#ลบลิงก์ที่นำไปสู่เว็บไซต์
def clean_url(maill):
  maill = str(maill)
  maill = maill.replace(' - ', '-')
  maill = maill.replace(':', '')
  maill = maill.replace('  / / ', '://')
  maill = maill.replace(' . ', '.')
  maill = maill.replace(' - ', '-')
  maill = maill.replace(' / ', '/')
  maill = maill.replace(' /', '/')
  maill = maill.replace(' ? ', '?')
  maill = maill.replace(' = ', '=')
  maill = maill.replace(' @ ', '@')
  maill = re.sub(r'http\S+', '', maill)
  return maill
mail_data['CleanedText'] = mail_data['CleanedText'].apply(clean_url)
```

ภาคผนวก ค.9 ชุดคำสั่งไพทอนในการลบตัวอักขระพิเศษ

```
#ลบอักขระพิเศษ
def remove_punctuations(maill):
    for punctuation in string.punctuation:
      maill = maill.replace(punctuation, ' ')
    return maill
mail_data['CleanedText'] = mail_data['CleanedText'].apply(remove_punctuations)
#ลบตัวเลข
def clean_num(maill):
  maill = str(maill)
  maill = re.sub('\d+', '', text)
  maill = ' '.join(maill.split())
  return maill
mail_data['CleanedText'] = mail_data['CleanedText'].apply(clean_num)
```

ภาคผนวก ค.10 ชุดคำสั่งไพทอนในการตัดข้อความออกเป็นคำ

```
#ตัดข้อความออกเป็นคำ
text_corpus = [ ]
for line in mail_data['CleanedText']:
    words = line.split(" ")
    text_corpus.append(words)
```

ภาคผนวก ค.11 ชุดคำสั่งไพทอนในการสร้างเวกเตอร์ของคำด้วย Word2Vec

```
#แปลงคำเป็นตัวเลขโดยใช้โมเดล Word2Vec
from gensim.models import Word2Vec, FastText, KeyedVectors
model = Word2Vec(text_corpus, min_count=0, size=100, workers = 4 )
wordstxt = list(model.wv.vocab)
```

ภาคผนวก ค.12 ชุดคำสั่งไพทอนในการลดมิติเวกเตอร์ของคำโดยใช้ PCA

```
#ลดมิติของข้อมูลโดยใช้ PCA
from sklearn.decomposition import PCA
newvec = model.wv[model.wv.vocab]
#กำหนดให้มิติของข้อมูลเป็น 2
pca = PCA(n_components=2)
result = pca.fit_transform(newvec)
#เก็บค่าเวกเตอร์ไว้ในตารางใหม่ชื่อ pca_df
pca_df = pd.DataFrame(result, columns = ['x','y'])
pca_df['Word'] = wordstxt
```

ภาคผนวก ค.13 ชุดคำสั่งไพทอนในการกำจัดค่านอกเกณฑ์ด้วยวิธี DBSCAN

```
#สร้างแบบจำลองDBSCAN
from sklearn.cluster import DBSCAN
x_y = pca_df[['x', 'y']]
db_input = x_y.to_numpy()
dbscan_cluster_model = DBSCAN(eps = 1, min_samples = 5).fit(db_input)
#สร้างตารางแสดงผลลัพท์ของแบบจำลองDBSCAN
pca_df['DB_cluster'] = dbscan_cluster_model.labels_
pca_df
pca_df['DB_cluster'].value_counts()
```

ภาคผนวก ค.14 ชุดคำสั่งไพทอนในการกำจัดค่านอกเกณฑ์ด้วยวิธีป่าไม้โดดเดี่ยว (Isolation Forest)

```
#สร้างแบบจำลองป่าไม้โดดเดี่ยว
from sklearn.ensemble import IsolationForest
model = IsolationForest(n_estimators=1000,max_samples='auto',contamination=float(
0.2),max_features=1.0)
model.fit(pca_df[['x']])
#สร้างตารางแสดงผลลัพท์ของแบบจำลองป่าไม้โดดเดี่ยว
pca_df['anomalies_score'] = model.decision_function(pca_df[['x']])
pca_df['anomaly'] = model.predict(pca_df[['x']])
pca_df.head(20)
```

ภาคผนวก ค.15 ชุดคำสั่งไพทอนในการสร้างแบบจำลอง IF-LOF

```
#สร้างแบบจำลอง IF-LOF
from sklearn.ensemble import IsolationForest
model =
IsolationForest(n_estimators=1000,max_samples='auto',contamination=0.02,max_feat
ures=1.0)
model.fit(anomaly_inputs)
df_if= pca_df.loc[pca_df['anomaly']== -1]
from sklearn.neighbors import LocalOutlierFactor
from sklearn.datasets import make_blobs
from numpy import quantile, where, random
import matplotlib.pyplot as plt
df_lf = pd.DataFrame(df_if, columns = ['x','y'])
lof = LocalOutlierFactor(n_neighbors=20, contamination=0.05)
y_pred = lof.fit_predict(df_lf)
#สร้างตารางแสดงผลลัพท์ของแบบจำลองป่าไม้โดดเดี่ยว
df_if['lof']=y_pred
```

ภาคผนวก ค.16 ชุดคำสั่งไพทอนในการกำจัดคำที่เป็นค่านอกเกณฑ์แต่ละวิธี

```
#กำจัดคำที่เป็นค่านอกเกณฑ์
for u in range(len(mail_data['(DB,IF,lof)'])):
  list_of_words = mail_data['(DB,IF,lof)'][u].split()
  for e in range(len(y['Word'])):
    list_of_words = [word for word in list_of_words if word.lower()
                not in y['Word'].iloc[e]]
  list_of_words = ' '.join(list_of_words)
  mail_data['(DB,IF,lof)'][u] = list_of_words
```

ภาคผนวก ค.17 ชุดคำสั่งไพทอนในการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีนาอีฟเบส์

```
#นำเข้าข้อมูลที่ไม่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์
input_data = pd.DataFrame({'text':data['CleanedText'], 'label':data['label']})
input_data.head()
x = input_data['text']
y = input_data['label']
#นำเข้าข้อมูลที่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์ของวิธี DBSCAN
input_data = pd.DataFrame({'text':mail_data['DB'], 'label':mail_data['label']})
input_data.head()
x = input_data['text']
y = input_data['label']
#นำเข้าข้อมูลที่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์ของวิธีป่าไม้โดดเดี่ยว
input_data = pd.DataFrame({'text':mail_data['IF'], 'label':mail_data['label']})
input_data.head()
x = input_data['text']
y = input_data['label']
#นำเข้าข้อมูลที่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์ของวิธี IF-LOF
input_data = pd.DataFrame({'text':mail_data['lof'], 'label':mail_data['label']})
input_data.head()
x = input_data['text']
y = input_data['label']
#แบ่งข้อมูล
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,stratify=y,test_size=0.3,random_state=
0)
#แปลงชุดข้อมูลฝึกสอนเป็นตัวเลข
from sklearn.feature_extraction.text import CountVectorizer
Vectorizer = CountVectorizer()
count= Vectorizer.fit_transform(x_train.values)
```

```
#สร้างแบบจำลองนาอีฟเบส์
from sklearn.naive_bayes import MultinomialNB
NB = MultinomialNB()
NB.fit(count, y_train)
```

ภาคผนวก ค.18 ชุดคำสั่งไพทอนในการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน

```
#นำเข้าข้อมูลที่ไม่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์
input_data = pd.DataFrame({'text':data['CleanedText'], 'label':data['label']})
input_data.head()
x = input_data['text']
y = input_data['label']
#นำเข้าข้อมูลที่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์ของวิธี DBSCAN
input_data = pd.DataFrame({'text':mail_data['DB'], 'label':mail_data['label']})
input_data.head()
x = input_data['text']
y = input_data['label']
#นำเข้าข้อมูลที่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์ของวิธีป่าไม้โดดเดี่ยว
input_data = pd.DataFrame({'text':mail_data['IF'], 'label':mail_data['label']})
input_data.head()
x = input_data['text']
y = input_data['label']
```

```
#แบ่งข้อมูล
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,stratify=y,test_size=0.3,random_state=
0)
#แปลงชุดข้อมูลฝึกสอนเป็นตัวเลข
from sklearn.feature_extraction.text import CountVectorizer
Vectorizer = CountVectorizer()
count= Vectorizer.fit_transform(x_train.values)
#สร้างแบบจำลองวิธีซัพพอร์ตเวกเตอร์แมชชีน
from sklearn import svm
SVM = svm.SVC(kernel='rbf')
SVM.fit(count, y_train)
```

ภาคผนวก ค.19 ชุดคำสั่งไพทอนในการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีเพื่อนบ้านใกล้สุด kตัว

```
#นำเข้าข้อมูลที่ไม่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์
input_data = pd.DataFrame({'text':data['CleanedText'], 'label':data['label']})
input_data.head()
x = input_data['text']
y = input_data['label']
#นำเข้าข้อมูลที่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์ของวิธี DBSCAN
input_data = pd.DataFrame({'text':mail_data['DB'], 'label':mail_data['label']})
input_data.head()
x = input_data['text']
y = input_data['label']
#นำเข้าข้อมูลที่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์ของวิธีป่าไม้โดดเดี่ยว
input_data = pd.DataFrame({'text':mail_data['IF'], 'label':mail_data['label']})
input_data.head()
x = input_data['text']
y = input  data['label']
```

```
#แบ่งข้อมูล
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,stratify=y,test_size=0.3,random_state=
0)
#แปลงชุดข้อมูลฝึกสอนเป็นตัวเลข
from sklearn.feature_extraction.text import CountVectorizer
Vectorizer = CountVectorizer()
count= Vectorizer.fit_transform(x_train.values)
#สร้างแบบจำลองเพื่อนบ้านใกล้สุด k ตัว
from sklearn.neighbors import KNeighborsClassifier
KNN = KNeighborsClassifier(n_neighbors = 1)
KNN.fit(count, y_train)
```

ภาคผนวก ค.20 ชุดคำสั่งไพทอนในการสร้างการเรียนรู้แบบรวมกลุ่มของการเรียนรู้ของเครื่อง

```
# ติดตั้งชุดคำสั่ง
from sklearn.metrics import log_loss
from sklearn.ensemble import VotingClassifier
# สร้างแบบจำลองการเรียนรู้ของเครื่อง
model_NB = MultinomialNB()
model_SVM = svm.SVC(kernel='rbf')
model_KNN = KNeighborsClassifier(n_neighbors = 1)
# สร้างแบบจำลองการเรียนรู้แบบรวมกลุ่มวิธีการโหวตเสียงข้างมาก
final_model = VotingClassifier(estimators=[('lr', model_NB), ('xgb', model_SVM), ('rf',
model_ KNN)], voting='hard')
final_model.fit(X_train, y_train)
pred_final = final_model.predict(X_test)
```

ภาคผนวก ค.21 ชุดคำสั่งไพทอนในการสร้างแบบจำลองการเรียนรู้เชิงลึกด้วยวิธีโครงข่ายประสาทเทียม

```
#นำเข้าข้อมูลที่ไม่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์
input_data = pd.DataFrame({'text':data['CleanedText'], 'label':data['label']})
input_data.head()
x = input_data['text']
y = input_data['label']
#นำเข้าข้อมูลที่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์ของวิธี DBSCAN
input_data = pd.DataFrame({'text':data['DB'], 'label':data['label']})
input_data.head()
x = input_data['text']
y = input_data['label']
#นำเข้าข้อมูลที่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์ของวิธีป่าไม้โดดเดี่ยว
input_data = pd.DataFrame({'text':mail_data['IF'], 'label':mail_data['label']})
input_data.head()
x = input_data['text']
y = input_data['label']
#นำเข้าข้อมูลที่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์ของวิธี IF-LOF
input_data = pd.DataFrame({'text':mail_data['lof'], 'label':mail_data['label']})
input_data.head()
x = input_data['text']
y = input_data['label']
#แบ่งข้อมูล
x_train, x_test, y_train, y_test = train_test_split(x, y,stratify=y,
test_size=0.3,random_state=0)
#นำชุดข้อมูลฝึกสอนมาหาขนาดของคำ
from keras.preprocessing.text import Tokenizer
tokenizer = Tokenizer()
tokenizer.fit_on_texts(x_train)
word_index = tokenizer.word_index
```

```
vocab_size = len(tokenizer.word_index) + 1000
#ปรับข้อมูลให้มีขนาดเท่ากัน
from tensorflow.keras.preprocessing.sequence import pad_sequences
X_train = pad_sequences(tokenizer.texts_to_sequences(x_train) , maxlen =
3338,2588,2420,3235)
X_test = pad_sequences(tokenizer.texts_to_sequences(x_test) , maxlen =
3338,2588,2420,3235)
#สร้างแบบจำลองโครงข่ายประสาทเทียม
from keras.layers import Dense, Embedding, Dropout
from keras.models import Sequential
ANN_model = Sequential()
ANN_model.add (Embedding(vocab_size,100,input_length=max_length))
ANN_model.add(Flatten())
ANN_model.add(Dense(32, activation='relu', input_dim=3338))
ANN_model.add(Dropout(0.5))
ANN_model.add(Dense(1, activation='sigmoid'))
ANN_model.compile(loss='binary_crossentropy', optimizer='adam',
metrics=['accuracy'])
ANN_model.fit(X_train, y_train, batch_size=100, epochs=100)
```

ภาคผนวก ค.22 ชุดคำสั่งไพทอนในการสร้างแบบจำลองการเรียนรู้เชิงลึกด้วยวิธีโครงข่ายประสาทแบบวนซ้ำ

```
#นำเข้าข้อมูลที่ไม่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์
input_data = pd.DataFrame({'text':data['CleanedText'], 'label':data['label']})
input_data.head()
x = input_data['text']
y = input_data['label']
#นำเข้าข้อมูลที่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์ของวิธี DBSCAN
input_data = pd.DataFrame({'text':data['DB'], 'label':data['label']})
input_data.head()
x = input_data['text']
y = input_data['label']
#นำเข้าข้อมูลที่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์ของวิธีป่าไม้โดดเดี่ยว
input_data = pd.DataFrame({'text':mail_data['IF'], 'label':mail_data['label']})
input_data.head()
x = input_data['text']
y = input_data['label']
#นำเข้าข้อมูลที่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์ของวิธี IF-LOF
input_data = pd.DataFrame({'text':mail_data['lof'], 'label':mail_data['label']})
input_data.head()
x = input_data['text']
y = input_data['label']
#แบ่งข้อมูล
x_train, x_test, y_train, y_test = train_test_split(x, y,stratify=y,
test_size=0.3,random_state=0)
#นำชุดข้อมูลฝึกสอนมาหาขนาดของคำ
from keras.preprocessing.text import Tokenizer
tokenizer = Tokenizer()
tokenizer.fit_on_texts(x_train)
word_index = tokenizer.word_index
```

```
vocab_size = len(tokenizer.word_index) + 1000
#ปรับข้อมูลให้มีขนาดเท่ากัน
from tensorflow.keras.preprocessing.sequence import pad_sequences
X_train = pad_sequences(tokenizer.texts_to_sequences(x_train) , maxlen =
3338,2588,2420,3235)
X_test = pad_sequences(tokenizer.texts_to_sequences(x_test) , maxlen =
3338,2588,2420,3235)
#สร้างแบบจำลองโครงข่ายประสาทแบบวนซ้ำ
from keras.layers import SimpleRNN, Embedding, Dropout
from keras.models import Sequential
RNN_model = Sequential()
RNN_model.add (Embedding(vocab_size,100,input_length=max_length))
RNN_model.add(SimpleRNN(32))
RNN_model.add(Dropout(0.5))
RNN_model.add(Dense(1, activation='sigmoid'))
RNN_model.compile(loss='binary_crossentropy', optimizer='adam',
metrics=['accuracy'])
RNN_model.fit(X_train, y_train, batch_size=100, epochs=100)
```

ภาคผนวก ค.23 ชุดคำสั่งไพทอนในการสร้างแบบจำลองการเรียนรู้เชิงลึกด้วยวิธีหน่วยความจำระยะสั้นยาว

```
#นำเข้าข้อมูลที่ไม่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์
input_data = pd.DataFrame({'text':data['CleanedText'], 'label':data['label']})
input_data.head()
x = input_data['text']
y = input_data['label']
#นำเข้าข้อมูลที่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์ของวิธี DBSCAN
input_data = pd.DataFrame({'text':data['DB'], 'label':data['label']})
input_data.head()
x = input_data['text']
y = input_data['label']
#นำเข้าข้อมูลที่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์ของวิธีป่าไม้โดดเดี่ยว
input_data = pd.DataFrame({'text':mail_data['IF'], 'label':mail_data['label']})
input_data.head()
x = input_data['text']
y = input_data['label']
#นำเข้าข้อมูลที่ผ่านการกำจัดคำที่เป็นค่านอกเกณฑ์ของวิธี IF-LOF
input_data = pd.DataFrame({'text':mail_data['lof'], 'label':mail_data['label']})
input_data.head()
x = input_data['text']
y = input_data['label']
#แบ่งข้อมูล
x_train, x_test, y_train, y_test = train_test_split(x, y,stratify=y,
test_size=0.3,random_state=0)
#นำชุดข้อมูลฝึกสอนมาหาขนาดของคำ
from keras.preprocessing.text import Tokenizer
tokenizer = Tokenizer()
tokenizer.fit_on_texts(x_train)
word_index = tokenizer.word_index
```

```
vocab_size = len(tokenizer.word_index) + 1000
#ปรับข้อมูลให้มีขนาดเท่ากัน
from tensorflow.keras.preprocessing.sequence import pad_sequences
X_train = pad_sequences(tokenizer.texts_to_sequences(x_train) , maxlen =
3338,2588,2420,3235)
X_test = pad_sequences(tokenizer.texts_to_sequences(x_test) , maxlen =
3338,2588,2420,3235)
#สร้างแบบจำลองหน่วยความจำระยะสั้นยาว
from keras.layers import LSTM, Embedding, Dropout
from keras.models import Sequential
lstm_model = Sequential()
lstm_model.add(Embedding(vocab_size,100,input_length=max_length))
lstm_model.add(LSTM(100))
lstm_model.add(Dropout(0.4))
lstm_model.add(Dense(16, activation= 'relu'))
lstm_model.add(Dropout(0.2))
lstm_model.add(Dense(1, activation= 'sigmoid'))
lstm_model.compile(loss='binary_crossentropy', optimizer='adam',
metrics=['accuracy'])
lstm_model.fit(X_train, y_train, batch_size=10, epochs=10)
```

ภาคผนวก ค.24 ชุดคำสั่งไพทอนในการสร้างการเรียนรู้แบบรวมกลุ่มของการเรียนรู้เชิงลึก

```
# สร้างแบบจำลองการเรียนรู้เชิงลึก
def ANN() :
    ANN_model = Sequential()
    ANN_model.add(Embedding(vocab_size,100,input_length=max_length))
    ANN_model.add(Flatten())
    ANN_model.add(Dense(32, activation='relu', input_dim=3338))
    ANN_model.add(Dropout(0.5))
    ANN_model.add(Dense(1, activation='sigmoid'))
    return ANN_model
def RNN() :
    RNN_model = Sequential()
    RNN_model.add(Embedding(vocab_size,100,input_length=max_length))
    RNN_model.add(SimpleRNN(32))
    RNN_model.add(Dropout(0.5))
    RNN_model.add(Dense(1, activation='sigmoid'))
    return RNN_model
def lstm() :
    lstm_model = Sequential()
    lstm_model.add(Embedding(vocab_size,100,input_length=max_length))
    lstm_model.add(LSTM(100))
    lstm_model.add(Dropout(0.4))
    lstm_model.add(Dense(16, activation= 'relu'))
    lstm_model.add(Dropout(0.2))
    lstm_model.add(Dense(1, activation= 'sigmoid'))
    return lstm_model
# สร้างแบบจำลองการสร้างการเรียนรู้แบบรวมกลุ่มของการเรียนรู้เชิงลึก
M1=ANN()
M2=RNN()
M3=lstm()
```

```
M1.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
M2.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
M3.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
M1.fit(X_train, y_train, batch_size=100, epochs=100)
M2.fit(X_train, y_train, batch_size=100, epochs=100)
M3.fit(X_train, y_train, batch_size=10, epochs=10)
y_predM1 = M1.predict(X_test)
y_predM2 = M2.predict(X_test)
y_predM3 = M3.predict(X_test)
y_pred_ann = np.round(y_predM1).flatten().astype(int)
y_pred_rnn = np.round(y_predM2).flatten().astype(int)
y_pred_lstm = np.round(y_predM3).flatten().astype(int)
y_pred_ensemble = (y_pred_lstm + y_pred_rnn+y_pred_ann)/3
for i in range(len(y_pred_ensemble)):
  if y_pred_ensemble[i]>0.5:
    y_pred_ensemble[i]=1
  else:
    y_pred_ensemble[i]=0
```

ภาคผนวก ค.25 ชุดคำสั่งไพทอนในการประเมินประสิทธิภาพของแบบจำลอง

```
#การนำชุดข้อมูลทดสอบไปเข้าแบบจำลอง
y_predict = model.predict(x_test)
#การประเมินประสิทธิภาพของแบบจำลอง
from sklearn import metrics
from sklearn.metrics import confusion_matrix
print("Accuracy:",metrics.accuracy_score(y_test, y_predict))
print("Precision:",metrics.precision_score(y_test, y_predict))
print("Recall:",metrics.recall_score(y_test, y_predict))
print("f1_score:",metrics.f1_score(y_test, y_predict))
```

งานทะเบียนคณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

คำรับรองเล่มปัญหาพิเศษ

วันที่ 24 เดือน พฤษภาคม พ.ศ. 2566

ข้าพเจ้า นายจิรายุ พิทักษ์ตันสกุล รหัสนักศึกษา 62050758

นายดิษย์ฉัตร เปียลาวัณ รหัสนักศึกษา 62050772

นางสาวปีย์วรานุช เบญญาบุณาพจน์ รหัสนักศึกษา 62050801

นักศึกษาหลักสูตรวิทยาศาสตรบัณฑิต สาขาวิชา สถิติประยุกต์ ภาควิชา สถิติ

ขอรับรองว่าปัญหาพิเศษ เรื่อง

การเปรียบเทียบประสิทธิภาพวิธีการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกใน
การตรวจจับจดหมายอิเล็กทรอนิกส์ไม่พึงประสงค์

EFFICIENCY COMPARISONS OF MACHINE LEARNING AND DEEP
LEARNING METHODS IN SPAM MAIL DETECTION

ปีการศึกษา 2565

เป็นผลงานวิจัยที่มิได้คัดลอกหรือละเมิดลิขสิทธิ์ของผู้อื่นและได้ผ่านการตรวจสอบความซ้ำซ้อนเรียบร้อยแล้ว และได้แนบเอกสารการตรวจสอบการลอกเลียนงานวรรณกรรมที่ตรวจสอบจากเล่มปัญหาพิเศษฉบับสมบูรณ์แล้ว

โปรแกรม Turnitin.....24....%

ลงชื่อ.....	ลงชื่อ.....	ลงชื่อ.....
(นายจิรายุ พิทักษ์ตันสกุล)	(นายดิษย์ฉัตร เปียลาวัณ)	(นางสาวปีย์วรานุช เบญญาบุณาพจน์)
นักศึกษา	นักศึกษา	นักศึกษา

ข้าพเจ้า รศ.สายชล สินสมบูรณ์ทอง อาจารย์ที่ปรึกษาปัญหาพิเศษ ได้ตรวจสอบปัญหาพิเศษของนักศึกษาข้างต้นแล้ว ขอรับรองว่าเป็นผลงานวิจัยของนักศึกษาจริงและมีเนื้อหาสมบูรณ์ จึงลงชื่อไว้เป็นหลักฐาน

ลงชื่อ...

(รศ.สายชล สินสมบูรณ์ทอง)

อาจารย์ที่ปรึกษา