การตรวจจับสิ่งผิดปรกติของการเข้าใช้ระบบ AZURE ACTIVE DIRECTORY ด้วยเทคนิคการเรียนรู้ของเครื่อง

DETECTING LOGON ANOMALY OF AZURE ACTIVE DIRECTORY SERVICE USING MACHINE LEARNING TECHNIQUES

พุฒิธร ชลิตชัยยะ

สหกิจศึกษานี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรบัณฑิต (สถิติประยุกต์)
ภาควิชาสถิติ คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ปีการศึกษา 2565

DETECTING LOGON ANOMALY OF AZURE ACTIVE DIRECTORY SERVICE USING MACHINE LEARNING TECHNIQUES

PUTTHITORN CHALITCHAIYA

A COOPERATIVE EDUCATION 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

สหกิจศึกษา	การตรวจจับสิ่งผิดปรกติของการเข้าใช้ระบบ Azure Active Directory ด้วยเทคนิคการเรียนรู้ของเครื่อง Detecting Logon Anomaly of Azure Active Directory Service using Machine Learning Techniques
ชื่อนักศึกษา	นาย พุฒิธร ชลิตชัยยะ รหัสนักศึกษา 62050806
ปริญญา	วิทยาศาสตรบัณฑิต (สถิติประยุกต์)
ภาควิชา	สถิติ
ปีการศึกษา	2565
อาจารย์ที่ปรึกษา	ผศ.ดร.พรพิมล ชัยวุฒิศักดิ์

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

คณะกรรมการสอบ	ลายมือชื่อ
ดร.สกุณา ศรีอโนมัย ประธานกรรมการ	
คุณวิศิษฐ์ กิจชัยนุกูล กรรมการ	
ผศ.ดร.พรพิมล ชัยวุฒิศักดิ์ กรรมการและอาจารย์ที่ปรึกษา	

สหกิจศึกษา การตรวจจับความผิดปรกติของการเข้าใช้ระบบ Azure Active Directory ด้วยเทคนิคการเรียนรู้ของเครื่อง
ชื่อนักศึกษา นาย พุฒิธร ชลิตชัยยะ รหัสนักศึกษา 62050806
ปริญญา วิทยาศาสตรบัณฑิต (สถิติประยุกต์)
ภาควิชา สถิติ
คณะ วิทยาศาสตร์
มหาวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง (สจล.)
ปีการศึกษา 2565
อาจารย์ที่ปรึกษา ผศ.ดร.พรพิมล ชัยวุฒิศักดิ์

บทคัดย่อ

ปัจจุบันองค์กรต่าง ๆ ต้องเผชิญกับความเสี่ยงจากภัยคุกคามทางไซเบอร์มากขึ้น ดังนั้นการรักษาความมั่นคงปลอดภัยต่อภัยคุกคามทางไซเบอร์ จึงมีบทบาทที่สำคัญต่อองค์กรเป็นอย่างมาก หนึ่งในแนวทางสำหรับการรักษาความปลอดภัย และควบคุมความเสี่ยงของระบบสารสนเทศ คือ การตรวจจับสิ่งผิดปรกติ ซึ่งจะช่วยให้ธุรกิจสามารถระบุแนวโน้มที่ผิดปรกติและพฤติกรรมการฉ้อโกง โดยการศึกษาวิจัยในครั้งนี้มีวัตถุประสงค์เพื่อตรวจจับสิ่งผิดปรกติที่เกิดขึ้นจาก Log File ในระบบ Azure Active Directory ของบริษัทผลิตเครื่องดื่มแห่งหนึ่งโดยอาศัยวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง วิธีไอโซเลชั่นฟอเรส และวิธีไอเอฟ-แอลโอเอฟ จากนั้นวิเคราะห์คุณลักษณะของข้อมูลสิ่งผิดปรกติที่ได้โดยอาศัยวิธีการจัดกลุ่มด้วยเทคนิคเคมีน และการทดสอบแมนน์-วิตนีย์ เพื่ออธิบายรูปแบบการเข้ามาใช้งานในแต่ละช่วงเวลา ผลการศึกษาพบว่าสัดส่วนของสิ่งผิดปรกติที่ตรวจจับด้วยวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง วิธีไอโซเลชั่นฟอเรส และวิธีไอเอฟ-แอลโอเอฟ คิดเป็นร้อยละ 1.57, 1.86, 2.00 และ 0.19 ตามลำดับ ซึ่งรูปแบบของสิ่งผิดปรกติที่ตรวจพบสามารถแบ่งได้เป็น 2 กลุ่ม โดยกลุ่มที่ 1 ประกอบด้วยสิ่งผิดปรกติเป็นจำนวนมาก ซึ่งจำนวนครั้งการเข้ามาใช้งานไม่แตกต่างกันมากนัก ในขณะที่ กลุ่มที่ 2 ประกอบด้วยสิ่งผิดปรกติที่จำนวนครั้งการเข้ามาใช้งานที่มีค่าผิดแผกแตกต่างไปจากค่าในกลุ่มที่ 1 อย่างเห็นได้ชัดเจน นอกจากนี้ยังพบว่าทุกตัวแปรสามารถใช้ในการจัดกลุ่มความถี่ผิดปรกติการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลวได้อย่างมีนัยสำคัญทางสถิติ ในขณะที่มีเพียง 3 ตัวแปรเท่านั้นที่สามารถใช้ในการจัดกลุ่มความถี่ผิดปรกติการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จได้อย่างมีนัยสำคัญทางสถิติ โดยอาศัยการทดสอบแมนน์ -วิตนีย์

คำสำคัญ: การตรวจจับสิ่งผิดปรกติ, การเรียนรู้ของเครื่อง, การจัดกลุ่มข้อมูล, การทดสอบแมนน์ - วิตนีย์

Title	Detecting Logon Anomaly of Azure Active Directory Service using Machine Learning Techniques
Students	Mr. Putthitorn Chalitchaiya Student ID 62050806
Degree	Bachelor of Science (Applied Statistics)
Department	Statistics
School	Science
University	King Mongkut's Institute of Technology Ladkrabang (KMITL)
Academic Year	2022
Advisor	Asst Prof. Dr. Pornpimol Chaiwuttisak

Abstract

The organizations are increasingly exposed to cyber threats, making the maintenance of security against these threats crucial for the organization. One important guideline for securing and controlling the risks in an information system is the detection of anomalies. This helps businesses identify unusual trends and fraudulent behavior. The purpose of this research is to detect anomalies in the log files of a beverage company's Azure Active Directory system. The study utilizes various methods such as 1) LOF method 2) OC-SVM method 3) Isolation Forest method and 4) IF-LOF method There were 1.57, 1.86, 2.00, and 0.19 percent of the detected Anomaly respectively. The patterns of anomaly detected can be divided into 2 groups, Group 1 contains a large number of abnormalities. data is not much different, while the second group contains a little number of anomalies where data by different ranges of times is significantly different from the values in the first cluster. In addition, it was found that all variables can be used for the statistically significant clustering of the anomaly number of logging into the system with a failure status. While only 3 variables could be used to classify the anomaly frequency of visits in systems with a statistically significant success status. by the Mann-Whitney.

Keywords: Anomaly Detection, Machine Learning, Clustering, The Mann - Whitney Test

กิตติกรรมประกาศ

สหกิจศึกษาฉบับนี้สำเร็จลุล่วงได้ด้วยความอนุเคราะห์จาก ผศ.ดร.พรพิมล ชัยวุฒิศักดิ์ อาจารย์ที่ปรึกษาสหกิจศึกษาที่ได้ให้คำปรึกษา แนวคิด และข้อคิดเห็นต่าง ๆ อันเป็นประโยชน์อย่างยิ่งในการทำสหกิจศึกษา อีกทั้งยังช่วยแก้ปัญหาที่เกิดขึ้นระหว่างการดำเนินงานตลอดจนตรวจทานแก้ไขข้อบกพร่องต่าง ๆ จนสหกิจศึกษาฉบับนี้เสร็จสมบูรณ์ ผู้วิจัยซาบซึ้งในความอนุเคราะห์จากท่านอาจารย์ และกราบขอบพระคุณเป็นอย่างสูง ณ โอกาสนี้

ขอขอบพระคุณ ดร.สกุณา ศรีอโนมัย ที่ให้เกียรติเป็นประธานกรรมการสอบสหกิจ ให้คำปรึกษา คำแนะนำ และสละเวลาตรวจทานแก้ไขข้อบกพร่องต่าง ๆ ตลอดการทำเล่มสหกิจศึกษาจนสหกิจศึกษาฉบับนี้เสร็จสมบูรณ์

ขอขอบพระคุณ คุณวิศิษฐ์ กิจชัยนุกุล คุณธนกร ปวีณาภรณ์ ที่ให้การสนับสนุนในด้านการศึกษา และการทำสหกิจศึกษา อีกทั้งสละเวลาในการตรวจสอบ และให้คำปรึกษาเกี่ยวกับการทำงานตลอดการทำสหกิจศึกษา

ขอขอบพระคุณคณาจารย์ภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบังทุกท่าน ที่ได้มอบวิชาความรู้ และประสบการณ์ พร้อมทั้งให้คำปรึกษา คำแนะนำ และช่วยเหลือในเรื่องต่าง ๆ แก่ผู้วิจัยตลอดมา

ขอขอบพระคุณบิดา มารดา และครอบครัวที่สนับสนุนในด้านการศึกษาเล่าเรียน ตลอดจนคอยช่วยเหลือ และให้กำลังใจผู้วิจัยเสมอมา รวมถึงขอบคุณเพื่อนนักศึกษา และพี่ร่วมงานที่ให้คำปรึกษา ให้กำลังใจในการทำสหกิจศึกษาครั้งนี้ สุดท้ายนี้ขอขอบคุณบริษัทกรณีศึกษาที่อนุญาตให้นำข้อมูลมาใช้งานในการทำสหกิจศึกษาฉบับนี้

พุฒิธร ชลิตชัยยะ

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญรูป

สารบัญรูป (ต่อ)

รูปที่ **หน้า**

สารบัญรูป (ต่อ)

รูปที่ **หน้า**

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญ

เทคโนโลยีและระบบสารสนเทศกลายเป็นเครื่องมือสำคัญต่อการขับเคลื่อนการดำเนินการธุรกิจขององค์กรให้มีความก้าวหน้าและรวดเร็ว ทำให้องค์กรต่าง ๆ ต้องปรับเปลี่ยนธุรกิจให้เข้าสู่ระบบดิจิทัล (Digital Transformation) ผลที่เกิดตามมาคือ องค์กรเหล่านั้นต้องเผชิญกับความเสี่ยงจากภัยคุกคามทางไซเบอร์ (Cyber Threats) ที่มากขึ้น ยกตัวอย่างเช่น การโจมตีแบบฟิชชิง (Phishing) เป็นการโจมตีทางไซเบอร์ที่เกิดขึ้นในองค์กรมากกว่าร้อยละ 80 อีกทั้งในกรณีที่บริษัทต่าง ๆ ถูกโจมตีจากไวรัสเรียกค่าไถ่ (Ransomware) ทำให้บริษัทต้องหยุดทำงานและเกิดการสูญเสียเงินประมาณ 8,500 เหรียญสหรัฐต่อชั่วโมง ส่งผลให้บริษัทกว่า 60% ปิดตัวลงภายในครึ่งปี เนื่องจากภัยคุกคามทางไซเบอร์ (Bitdefender, 2021) จากการสำรวจความปลอดภัยทางไซเบอร์ของประเทศไทยในปี 2565 (ณิชาภา, 2565) พบว่าในระยะเวลา 12 เดือน ภัยคุกคามทางไซเบอร์ ที่สร้างความเสียหายหรือผลกระทบต่อหน่วยงานส่วนใหญ่มากที่สุด 4 อันดับแรก ได้แก่ ภัยคุกคามที่เกิดจากโปรแกรมซอฟต์แวร์ที่ถูกพัฒนาขึ้นเพื่อส่งผลลัพธ์ที่ไม่พึงประสงค์กับผู้ใช้งานหรือระบบ (Malicious Code) คิดเป็นร้อยละ 54 ภัยคุกคามที่เกิดจากการโจมตีสภาพความพร้อมใช้งานของระบบ (Availability) คิดเป็นร้อยละ 18 การพยายามเก็บข้อมูลเป้าหมาย (Information Gathering) คิดเป็นร้อยละ 16 และเหตุจากการถูกบุกรุกหรือเจาะระบบ (Intrusions Attempt) คิดเป็นร้อยละ 12 ดังนั้นการรักษาความปลอดภัยทางไซเบอร์จึงมีบทบาทที่สำคัญต่อธุรกิจและองค์กรเป็นอย่างมาก (ศุภโชคชัย, 2564) ดังนั้นการตรวจจับภัยคุกคามทางไซเบอร์จึงกลายเป็นประเด็นที่ได้รับความสนใจในปัจจุบัน

การตรวจจับภัยคุกคามทางไซเบอร์สามารถทำได้ 2 แนวทางคือ การตรวจจับการใช้ในทางที่ผิด (Misuse Detection) และการตรวจจับสิ่งผิดปรกติ (Anomaly Detection) โดยการตรวจจับการใช้ในทางที่ผิดจะสามารถตรวจสอบได้เฉพาะการโจมตีที่ทราบอยู่แล้วในระบบฐานข้อมูล โดยการกำหนดกฎหรือตัวกรองในระบบตรวจสอบ แต่ในขณะที่วิธีของการตรวจจับสิ่งผิดปรกตินั้นเป็นการตรวจสอบการบุกรุกระบบ โดยอาศัยหลักการที่ว่าพฤติกรรมการบุกรุก คือ พฤติกรรมที่มีลักษณะเบี่ยงเบนไปจากลักษณะการใช้งานทั่วไปในสภาวะที่ปรกติ โดยงานวิจัยของธนชัย (2559) ได้ทำการตรวจจับภัยคุกคามทางไซเบอร์จาก Log File ด้วยเทคนิคการตรวจจับสิ่งผิดปรกติ ซึ่งเป็นวิธีการที่สะดวก และนิยมใช้เพื่อที่จะทำให้ค้นพบรูปแบบ หรือภัยคุกคามที่ไม่เคยพบมาก่อน

วิธีการตรวจจับสิ่งผิดปรกติสามารถทำได้โดยใช้เทคนิคการเรียนรู้แบบมีผู้สอน (Supervised Learning) และการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) แต่เนื่องจากเทคนิคการเรียนรู้แบบมีผู้สอน (Supervised Learning) จำเป็นต้องใช้ป้ายกำกับคำตอบของข้อมูล (Label) จำนวนมากซึ่งมักจะเป็นไปไม่ได้ในสภาพแวดล้อมจริง ในส่วนของเทคนิคการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) เป็นการเรียนรู้ที่ไม่มีป้ายกำกับคำตอบของข้อมูล ได้แก่ วิธีโลคอลเอาท์-ไลเออร์แฟคเตอร์ (Local Outlier Factor: LOF) วิธีการจัดกลุ่ม (Clustering) วิธีซัพพอร์ตเวกเตอร์-แมชชีน (Support Vector Machine: SVM) วิธีการเรียนรู้แบบรวมกลุ่ม (Ensemble) (Jääskelä, 2020) ซึ่งเป็นวิธีที่ได้รับความนิยมในการตรวจจับสิ่งผิดปรกติ

ดังนั้นงานวิจัยนี้มุ่งศึกษาการตรวจจับสิ่งผิดปรกติจากแฟ้มข้อมูลเก็บบันทึกเหตุการณ์ที่เกิดขึ้น (Log file) ในระบบ Azure Active Directory ของบริษัทผลิตเครื่องดื่มแห่งหนึ่ง โดยวิธีการตรวจจับสิ่งผิดปรกติทั้งหมด 4 วิธี คือ วิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ (Local Outlier Factor: LOF) วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง (One-Class Support Vector Machine: OC-SVM) วิธีไอโซเลชั่นฟอเรส (Isolation Forest: IF) และวิธีไอเอฟ-แอลโอเอฟ (Isolation Forest- Local Outlier Factor: IF-LOF) จากนั้นวิเคราะห์สิ่งผิดปรกติที่ได้โดยอาศัยวิธีการจัดกลุ่ม (Clustering) ด้วยเทคนิควิธี K-Means Clustering และการทดสอบแมนน์ -วิตนีย์ (The Mann – Whitney test) เพื่ออธิบายพฤติกรรม และคุณลักษณะสิ่งที่ผิดปรกติของการเข้ามาใช้งานระบบในแต่ละช่วงเวลา

1.2 วัตถุประสงค์

1) เพื่อศึกษาวิธีการตรวจจับสิ่งผิดปรกติของการเข้ามาใช้งานในระบบด้วยเทคนิคการตรวจจับสิ่งผิดปรกติ (Anomaly Detection)
2) วิเคราะห์ผลการตรวจจับสิ่งผิดปรกติของการเข้ามาใช้งานในระบบโดยอาศัยวิธีการจัดกลุ่ม (Clustering) และสถิติเชิงอนุมาน

1.3 ขอบเขตของงานวิจัย

1. ขอบเขตด้านข้อมูล
   การศึกษาครั้งนี้ได้ดำเนินการศึกษากับบริษัทผลิตเครื่องดื่มแห่งหนึ่ง โดยทำการรวบรวมการเข้าใช้งานระบบ Azure Active Directory ของพนักงานบริษัทจาก Log File บนระบบ Microsoft Azure ซึ่งข้อมูลแบ่งออกเป็น 2 ชุด ดังนี้
   - ข้อมูลชุดที่ 1 ประกอบด้วยข้อมูล วันที่และเวลาที่ทำรายการ รหัสผู้ใช้งาน สถานะการเข้ามาใช้งานระบบ (สำเร็จ) จำนวน 1,000,000 รายการ
   - ข้อมูลชุดที่ 2 ประกอบด้วยข้อมูล วันที่และเวลาที่ทำรายการ รหัสผู้ใช้งาน สถานะการเข้ามาใช้งานระบบ (ล้มเหลว) จำนวน 253,005 รายการ
2. ขอบเขตด้านระยะเวลา
   ระยะเวลาของข้อมูลที่ใช้การศึกษาวิจัยเก็บรวบรวมตั้งแต่วันที่ 9 มกราคม พ.ศ. 2566 ถึง วันที่ 16 มีนาคม พ.ศ. 2566

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. เพื่อทราบถึงวิธีในการตรวจจับสิ่งผิดปรกติของการเข้ามาใช้งานในระบบ และวิเคราะห์คุณลักษณะของสิ่งผิดปรกติที่ได้ โดยอาศัยวิธีการจัดกลุ่มและสถิติเชิงอนุมาน
2. นำผลลัพธ์ของวิธีในการตรวจจับสิ่งผิดปรกติ ไปเป็นแนวทางประกอบในการพิจารณาสิ่งผิดปรกติของการเข้ามาใช้งานในระบบ

1.5 นิยามศัพท์

ในงานวิจัยครั้งนี้ ผู้วิจัยได้กำหนดความหมายของคำศัพท์ที่เกี่ยวข้องไว้ เพื่อทำให้ผู้ที่จะนำงานวิจัยนี้ไปศึกษาต่อได้เกิดความเข้าใจในแนวทางเดียวกัน ดังนี้

1) ความปลอดภัยทางไซเบอร์ (Cyber Security) คือ กระบวนการหรือการกระทำทั้งหมดที่จำเป็น เพื่อทำให้องค์กรปราศจากความเสี่ยง และความเสียหายที่มีผลต่อความปลอดภัยของข้อมูลข่าวสาร (Information) ในทุกรูปแบบ รวมถึงการระวังป้องกันต่อการอาชญากรรม การโจมตี การบ่อนทำลาย การโจรกรรม และความผิดพลาดต่าง ๆ โดยควรคำนึงถึงองค์ประกอบพื้นฐานของความปลอดภัยของข้อมูล หรือ CIA 3 ประการ ได้แก่ การรักษาความลับของข้อมูล (Confidentiality) การรักษาความคงสภาพของข้อมูลหรือความสมบูรณ์ของข้อมูล (Integrity) และความพร้อมใช้งานของข้อมูล (Availability) (สำนักงานรัฐบาลอิเล็กทรอนิกส์, 2559)
2) แฟ้มข้อมูลเก็บบันทึกเหตุการณ์ที่เกิดขึ้น (Log File) คือ แฟ้มข้อมูลเก็บบันทึกเหตุการณ์ที่เกิดขึ้นของข้อมูลจราจรคอมพิวเตอร์ ที่เกี่ยวข้องกับการติดต่อสื่อสารของระบบคอมพิวเตอร์ แสดงถึงแหล่งกำเนิด ต้นทาง ปลายทาง เส้นทาง เวลา วันที่ ปริมาณ ระยะเวลาชนิดของบริการ หรืออื่น ๆ ที่เกี่ยวข้องกับการติดต่อสื่อสารของระบบคอมพิวเตอร์ (สมชาย, 2566)
3) สิ่งผิดปรกติ (Anomaly) คือ รูปแบบที่เบี่ยงเบนอย่างมากจากพฤติกรรมที่คาดไว้ หรือพฤติกรรมที่เป็นปรกติของชุดข้อมูล สามารถกำหนดลักษณะเป็นจุดข้อมูลหรือรูปแบบที่แปลกแยก คาดไม่ถึง หรือผิดปรกติเมื่อเทียบกับข้อมูลส่วนใหญ่ (Hawkins, 1980)
4) การตรวจจับสิ่งผิดปรกติ (Anomaly Detection) คือการใช้วิธีทางคณิตศาสตร์ สถิติ, Data Mining, Machine Learning หรือ AI ดึงข้อมูลในอดีต (Historical Data) เพื่อคัดแยกสิ่งผิดปรกติ (Anomaly) และ ข้อมูลที่เป็นปรกติ (Normally) ออกจากกัน (ชาคริต, 2563)
5) การเรียนรู้ของเครื่อง (Machine Learning) คือ การทำให้ระบบคอมพิวเตอร์เรียนรู้และสร้างขั้นตอนวิธี (Algorithm) ที่สามารถเรียนรู้ข้อมูลและทำนายข้อมูลได้ (วีระพันธ์, 2564)
6) การเรียนรู้โดยไม่มีผู้สอน (Unsupervised Learning) คือ เป็นการเรียนรู้ที่ให้เครื่องจักรนั้นสามารถเรียนรู้ได้ด้วยตนเอง โดยไม่ต้องมีค่าเป้าหมายของแต่ละข้อมูล ซึ่งวิธีการคือมนุษย์จะเป็นผู้ใส่ข้อมูลต่าง ๆ และกำหนดสิ่งที่ต้องการจากข้อมูลเหล่านั้น โดยให้เครื่องจักรวิเคราะห์จากการจำแนกและสร้างแบบแผนจากข้อมูลที่ได้รับมา (พิพัฒน์, 2563)
7) วิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ (Local Outlier Factor: LOF) คือ การเปรียบเทียบความหนาแน่นของข้อมูลจุดต่างๆ แล้วแยกจุดที่มีความหนาแน่นน้อยออกเป็น สิ่งผิดปรกติ โดยความหนาแน่นจะคำนวณจาก K-Nearest neighbors ซึ่งก็คือระยะห่างระหว่างจุดที่เราสนใจ กับจุด "เพื่อนบ้าน" ที่ใกล้ที่สุดจำนวน K จุดตามที่ผู้วิจัยกำหนด (ชาคริต, 2563)
8) วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง (One-Class Support Vector Machine: OC-SVM) เป็นโมเดลที่ไม่มีการควบคุมสำหรับการตรวจจับสิ่งผิดปรกติ ซึ่งแตกต่างจากวิธีซัพพอร์ตเวกเตอร์แมชชีนที่ได้รับการดูแล วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่งไม่มีป้ายกำกับเป้าหมายสำหรับกระบวนการฝึกอบรมแบบจำลอง แต่จะเรียนรู้ขอบเขตของจุดข้อมูลปรกติและระบุข้อมูลที่อยู่นอกขอบเขตว่าเป็นความผิดปรกติแทน (Amy, 2021)

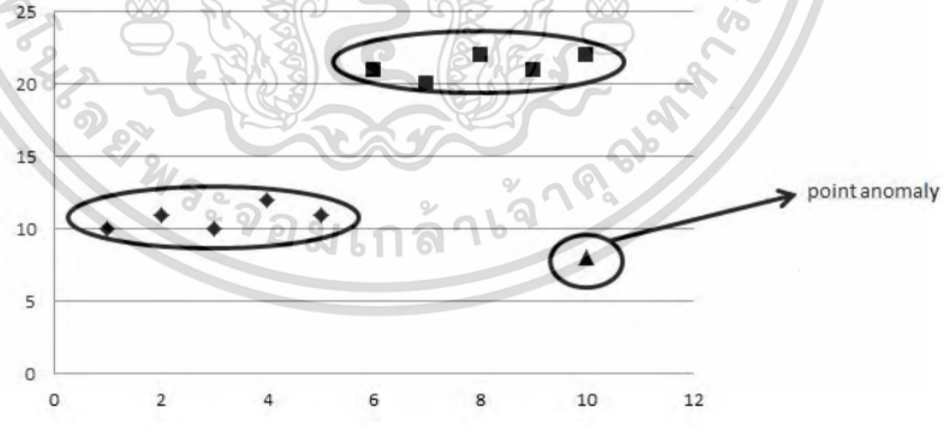
บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในการศึกษาวิจัยครั้งนี้ ผู้วิจัยได้มีการศึกษาแนวความคิดและทฤษฎีที่เกี่ยวข้อง ก่อนการทำการวิจัยเพื่อใช้ประกอบการศึกษางานวิจัยเรื่อง "การตรวจจับสิ่งผิดปกติของการเข้าใช้ระบบ Azure Active Directory ด้วยเทคนิคการเรียนรู้ของเครื่อง" โดยมีรายละเอียดดังต่อไปนี้

2.1 สิ่งผิดปกติ (Anomaly)

สิ่งผิดปกติ (Anomaly) คือ รูปแบบที่เบี่ยงเบนอย่างมากจากพฤติกรรมที่คาดไว้ หรือพฤติกรรมที่เป็นปกติของชุดข้อมูล (Hawkins, 1980) สามารถกำหนดลักษณะเป็นจุดข้อมูลหรือรูปแบบที่แปลกแยก คาดไม่ถึง หรือผิดปกติเมื่อเทียบกับข้อมูลส่วนใหญ่ สิ่งผิดปกติอาจเกิดขึ้นได้จากหลายสาเหตุ เช่น ข้อผิดพลาดในการรวบรวมข้อมูล มาตรวัดที่ไม่ถูกต้องของข้อมูล ค่าผิดปกติในการกระจายข้อมูล หรือตัวอย่างที่ผิดปกติจริงของข้อมูล การตรวจจับสิ่งผิดปกติมีความสำคัญในหลายด้าน เนื่องจากสามารถให้ข้อมูลเชิงลึก และยังสามารถระบุปัญหาหรือสิ่งผิดปกติที่อาจเกิดขึ้นและช่วยในกระบวนการตัดสินใจ ความผิดปกติอาจมีรูปแบบที่แตกต่างกันขึ้นอยู่กับลักษณะของข้อมูลและปัญหาที่เกิดขึ้น ประเภทของความผิดปกติ (Chandola et al, 2009) ได้แก่:

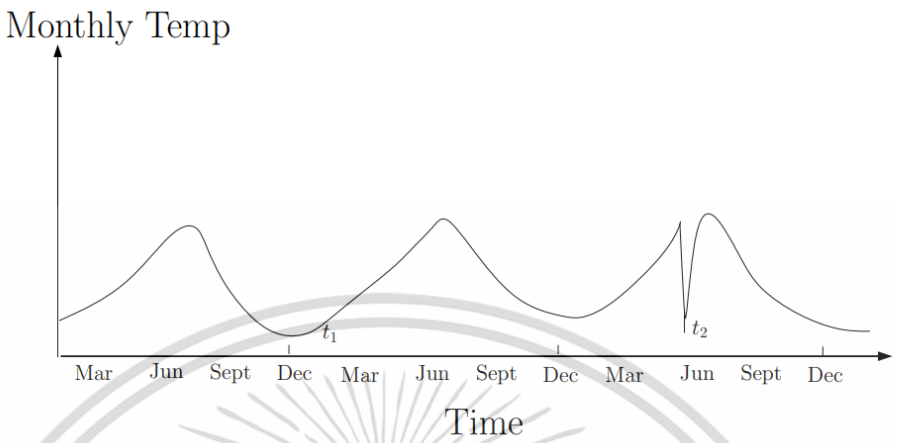
1. ความผิดปกติของจุด: จุดเหล่านี้คือตัวอย่างของข้อมูลแต่ละรายการที่มีความแตกต่างอย่างมากจากจุดข้อมูลส่วนใหญ่ ตัวอย่างเช่น ในชุดข้อมูลของอุณหภูมิ ความผิดปกติของจุดอาจเป็นค่าอุณหภูมิสูงสุดหรือต่ำสุด



รูปที่ 2.1 ตัวอย่างความผิดปกติของจุด
(ที่มา: Baddar et al., 2014)

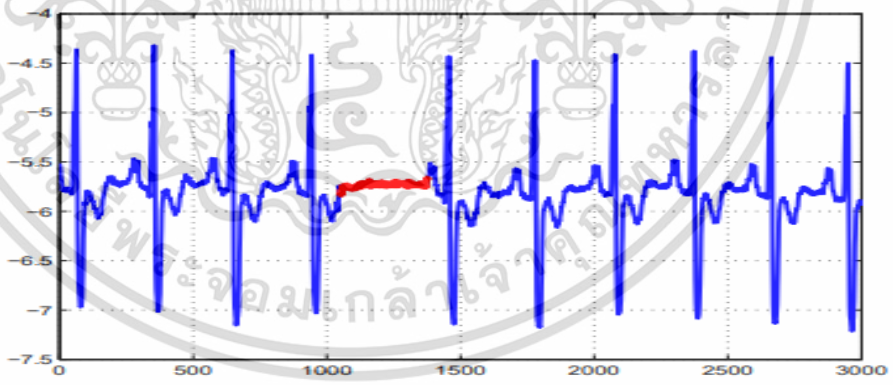
2. ความผิดปกติตามบริบท: ความผิดปกติเหล่านี้คือตัวอย่างของข้อมูลที่ถือว่าผิดปกติภายในบริบทหรือเงื่อนไขเฉพาะ ข้อมูลเหล่านี้อาจไม่ผิดปกติเมื่อพิจารณาแยกกัน แต่จะกลายเป็นสิ่งผิดปกติเมื่อพิจารณาภายในบริบท ตัวอย่างเช่น อุณหภูมิที่เวลา T1 และ

T2 จะเหมือนกันแต่เกิดในบริบทที่แตกต่างกัน ในกรณีนี้อุณหภูมิที่เวลา T2 จะถูกพิจารณาเป็นความผิดปรกติ



รูปที่ 2.2 ตัวอย่างความผิดปรกติตามบริบท
(ที่มา: Chandola et al., 2009)

3. ความผิดปรกติโดยรวม: หมายถึงกลุ่มของตัวอย่างข้อมูลที่แสดงพฤติกรรมที่ผิดปรกติเมื่อพิจารณาร่วมกัน แม้ว่าจุดข้อมูลแต่ละจุดภายในกลุ่มอาจไม่ผิดปรกติ แต่พฤติกรรมโดยรวมหรือความสัมพันธ์ระหว่างกันนั้นถือว่าผิดปรกติ ตัวอย่างเช่น คลื่นไฟฟ้าหัวใจของมนุษย์ บริเวณที่เน้นสีแดงแสดงถึงความผิดปรกติ เนื่องจากมีค่าต่ำเป็นเวลานานผิดปรกติ



รูปที่ 2.3 ตัวอย่างความผิดปรกติโดยรวม (ที่มา: Chandola et al., 2009)

2.2 แนวทางในการตรวจจับสิ่งผิดปรกติ

Patcha and Park (2007) กล่าวว่าในการตรวจจับสิ่งผิดปรกติมี 3 แนวทาง ได้แก่ การตรวจจับสิ่งผิดปรกติด้วยสถิติ (Statistical Anomaly Detection) การทำเหมืองข้อมูล (Data Mining Based Anomaly Detection) การเรียนรู้ของเครื่อง (Machine Learning Based Anomaly Detection) ดังต่อไปนี้

2.2.1 การตรวจจับสิ่งผิดปรกติด้วยสถิติ (Statistical Anomaly Detection)

การตรวจจับสิ่งผิดปรกติด้วยสถิติ (Statistical Anomaly Detection) เป็นเทคนิคหนึ่งที่ใช้ในการวิเคราะห์และตรวจสอบคุณสมบัติของข้อมูลเพื่อค้นหาสิ่งที่ผิดปรกติ วิธีการนี้ใช้ค่าทางสถิติที่กำหนดเองหรือค่าที่ได้จากข้อมูลในอดีตเป็นเกณฑ์ในการเปรียบเทียบและตรวจสอบกระจายของข้อมูลที่สนใจ ถ้าข้อมูลเบี่ยงเบนหรือแตกต่างจากค่าเกณฑ์ที่กำหนดจะถือว่ามีการเกิดสิ่งผิดปรกติขึ้น วิธีการนี้มีข้อดีหลายอย่าง เช่น ไม่ต้องมีความรู้ล่วงหน้าเกี่ยวกับข้อบกพร่องด้านความปลอดภัยหรือการโจมตี เพื่อให้ระบบสามารถตรวจจับ "ซีโร่เดย์" หรือการโจมตีล่าสุดได้ นอกจากนี้ วิธีการทางสถิติยังสามารถให้การแจ้งเตือนที่ถูกต้องเกี่ยวกับกิจกรรมที่เป็นอันตราย โดยมีตัวอย่างเช่นงานวิจัย Haystack (1988) เป็นระบบตรวจจับการบุกรุกตามความผิดปรกติทางสถิติแรกที่เกิดขึ้น ซึ่งใช้กลยุทธ์ในการตรวจจับความผิดปกติตามผู้ใช้และกลุ่มของข้อมูล และใช้พารามิเตอร์ของระบบแบบจำลองเป็นตัวแปรสุ่มแบบเกาส์เซียนอิสระ อย่างไรก็ตาม วิธีการนี้อาจมีข้อเสียอย่างเช่นหากใช้เกณฑ์ทางสถิติเป็นเพียงอย่างเดียว ระบบตรวจจับอาจถูกฝึกฝนให้มองเห็นสิ่งผิดปรกติเป็นสิ่งปรกติได้

2.2.2 การทำเหมืองข้อมูล (Data Mining Based Anomaly Detection)

การทำเหมืองข้อมูล (Data Mining) เป็นกระบวนการที่นำเสนอวิธีการวิเคราะห์ข้อมูลเพื่อค้นหารูปแบบ ความสัมพันธ์ หรือความเบี่ยงเบนที่อาจมองไม่เห็นด้วยตาเปล่า โดยการใช้เทคนิคเหมืองข้อมูล (Data Mining Techniques) เราสามารถค้นหาข้อมูลที่ซ่อนอยู่ในชุดข้อมูลขนาดใหญ่มักจะใช้กฎหรือรูปแบบที่กำหนดไว้ล่วงหน้าในข้อมูลเพื่อระบุสิ่งผิดปกติ ซึ่งมีข้อจำกัดคือใช้เวลานาน และจำเป็นต้องมีผู้เชี่ยวชาญในการระบุสิ่งผิดปรกติ (Expert Knowledge) ในการตรวจหาสิ่งผิดปรกติโดยใช้เหมืองข้อมูล (Data Mining-based Anomaly Detection) สามารถนำเสนอเทคนิคและขั้นตอนวิธีต่าง ๆ เพื่อค้นหาความผิดปกติในข้อมูลได้ หลักการทำงานของเหมืองข้อมูลที่ใช้สำหรับการตรวจหาสิ่งผิดปกติมักเน้นไปที่การค้นหาความแตกต่างหรือความเบี่ยงเบนออกมาจากข้อมูลปรกติ (Normal Data) โดยใช้เทคนิคต่าง ๆ เช่น Genetic Algorithms คือเทคนิคการคำนวณที่ใช้ลอจิกเสมือนกระบวนการวิวัฒนาการเพื่อแก้ปัญหาการค้นหาและปรับให้เหมาะสมในปัญหาที่ซับซ้อน โดยการสร้างประชากรของคำตอบเชิงพันธุกรรมและใช้กระบวนการเลือก การผสมพันธุ์ และการกลายพันธุ์เพื่อพัฒนาคำตอบที่ดีขึ้นตามเวลา และอีกเทคนิคคือ Association Rule Discovery เป็นกระบวนการทางคณิตศาสตร์และการเรียนรู้ของเครื่อง (Machine Learning) ซึ่งมุ่งเน้นการค้นหาความสัมพันธ์และความสัมพันธ์ที่น่าสนใจระหว่างรายการ (Items) ในชุดข้อมูล (Dataset) ซึ่งแสดงถึงความสัมพันธ์ทางสถิติซึ่งสามารถช่วยให้เราเข้าใจแนวโน้มและความสัมพันธ์ที่มีอยู่ในข้อมูล

2.2.3 การเรียนรู้ของเครื่อง (Machine Learning Based Anomaly Detection)

การเรียนรู้ของเครื่อง (Machine Learning) ในการตรวจจับความผิดปรกติ (Anomaly Detection) เป็นกระบวนการที่ใช้เทคนิคและวิธีการทางคอมพิวเตอร์ในการค้นหาและระบุค่าผิดปรกติหรือพฤติกรรมที่ไม่ปรกติในชุดข้อมูล เพื่อช่วยป้องกันการฉ้อโกง การโจมตี และการบุกรุกเครือข่าย ซึ่งการตรวจจับความผิดปรกติสามารถทำได้โดยใช้เทคนิคการเรียนรู้แบบมีผู้สอน (Supervised Learning) หรือแบบไม่มีผู้สอน (Unsupervised Learning) ขึ้นอยู่กับความสามารถ

และประสบการณ์ที่ต้องการในการดำเนินการตรวจจับความผิดปรกติของระบบในแต่ละกรณี ดังนั้นจึงมีวิธีการที่แตกต่างกันอย่างมากในการเรียนรู้และตรวจจับความผิดปรกติ ดังนี้

1. เทคนิคการเรียนรู้แบบมีผู้สอน (Supervised Learning) ในการตรวจจับความผิดปรกติแบบมีผู้สอน ต้องใช้ชุดข้อมูลการฝึกอบรมที่มีป้ายกำกับคำตอบ เช่น ข้อมูลที่แสดงถึงสภาวะปรกติและสภาวะผิดปรกติ โดยป้ายกำกับให้เป็นค่าบอกว่าแต่ละตัวอย่างในชุดข้อมูลเป็นปรกติหรือผิดปรกติ และใช้ข้อมูลเหล่านี้ในกระบวนการฝึกอบรมโมเดล เช่น แบบจำลองเชิงเส้น (Linear Models) หรือต้นไม้ตัดสินใจ (Decision Trees) เพื่อสร้างโมเดลที่สามารถรู้จักและจำแนกสิ่งที่เป็นปรกติและผิดปรกติได้ โดยในกระบวนการฝึกอบรม โมเดลจะเรียนรู้จากตัวอย่างที่มีป้ายกำกับและพยากรณ์ค่าผิดปกติของตัวอย่างใหม่ที่ไม่มีป้ายกำกับ

2. เทคนิคการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) ในการตรวจจับความผิดปรกติแบบไม่มีผู้สอน ไม่ต้องใช้ป้ายกำกับคำตอบในข้อมูลการฝึกอบรม เพียงแค่ใช้ข้อมูลและอัลกอริทึมเพื่อค้นหาโครงสร้างหรือลักษณะที่ไม่ปรกติภายในชุดข้อมูล เช่น การสร้างโมเดลการจัดกลุ่ม (Clustering Models) เพื่อระบุกลุ่มที่มีพฤติกรรมที่แตกต่างจากกลุ่มปรกติ หรือเทคนิคการประมวลผลสัญญาณ (Signal Processing Techniques) เพื่อตรวจจับความผิดปรกติทางสถิติในชุดข้อมูล

เทคนิคการเรียนรู้แบบมีผู้สอนและแบบไม่มีผู้สอนมีข้อดีและข้อเสียที่แตกต่างกัน การเรียนรู้แบบมีผู้สอนต้องการข้อมูลที่มีป้ายกำกับคำตอบ ซึ่งอาจจำเป็นต้องการความสอดคล้องและความถูกต้องของป้ายกำกับที่ส่งผลต่อประสิทธิภาพของโมเดล ในขณะที่การเรียนรู้แบบไม่มีผู้สอนมีความยืดหยุ่นในการปรับใช้กับชุดข้อมูลใหม่แต่มีความเชื่อถือน้อยกว่า ในบางกรณีอาจพบปัญหาในการระบุค่าผิดปรกติที่แม่นยำและการตรวจจับความผิดปรกติที่ไม่แม่นยำพอในชุดข้อมูลใหม่ อีกทั้งยังมีความซับซ้อนในการเลือกและปรับพารามิเตอร์ของอัลกอริทึมในกระบวนการเรียนรู้แบบไม่มีผู้สอน

2.3 การตรวจจับสิ่งผิดปรกติด้วยเทคนิคการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning)

เนื่องจากข้อมูลไม่มีป้ายกำกับคำตอบ ในการตรวจจับสิ่งผิดปรกติของข้อมูล จึงเป็นไปได้ยากหากต้องหาสิ่งผิดปรกติด้วยตนเอง จึงมีการนำวิธีการเรียนรู้ของเครื่องแบบไม่มีผู้สอนมาใช้เพื่อคัดแยกสิ่งผิดปรกติ (Anomaly) และสิ่งปรกติ (Normal) ออกจากกัน มีขั้นตอนวิธีดังนี้

2.3.1 วิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ (Local Outlier Factor: LOF)

Breunig et al. (2000) ได้เสนอวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์เป็นการแบ่งกลุ่มข้อมูลพิจารณาจากความหนาแน่นภายในกลุ่มข้อมูล โดยสร้างจากการค้นหาจำนวนเพื่อนบ้านใกล้ที่สุดที่เป็นพารามิเตอร์ K หรือ MinPts แนวคิด คือการเปรียบเทียบความหนาแน่นเฉลี่ยของเพื่อนบ้านที่ใกล้ที่สุดของจุดกับความหนาแน่นของเพื่อนบ้านภายในพื้นที่ วิธีโลคอลเอาท์ไลเออร์แฟคเตอร์มีขั้นตอนคำนวณดังนี้ (Goldstein and Uchida, 2016)

1) กำหนดค่าเพื่อนบ้านใกล้เคียง K (K – Nearest Neighbor) ในการคำนวณระยะห่างของเพื่อนบ้านใกล้เคียง ซึ่งหมายความว่าต้องหาเพื่อนบ้านที่ใกล้ที่สุดของแต่ละจุด
2) คำนวณหาระยะห่างมากที่สุดจากระยะห่างของเพื่อนบ้านใกล้เคียง ดังสมการที่ 2.1

$$reachdist_k(p,q) = max\{dist_k(q), d(p,q)\} \quad (2.1)$$

3) คำนวณหาจำนวนเพื่อนบ้านใกล้เคียง ดังสมการที่ 2.2

$$N_{dist_k(p)} = \{q \mid q \in D, d(p,q) \le dist_k(p)\} \quad (2.2)$$

4) คำนวณหาความหนาแน่นของแต่ละจุดนั้นประมาณจากการคำนวณความหนาแน่นของการเข้าถึง (Local reachability density : LRD) ดังสมการที่ 2.3

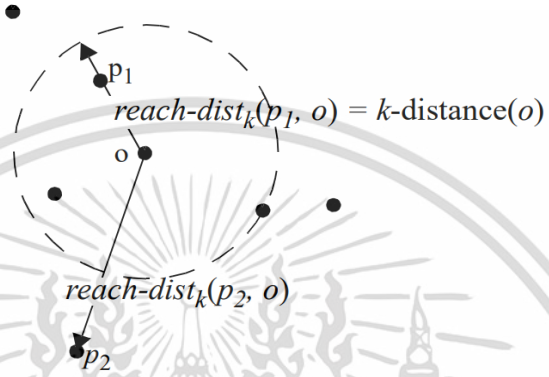
$$lrd_k(p) = \frac{1}{\left(\frac{\sum_{q \in N_k(p)} reach - dist_k(p,q)}{|N_k(p)|}\right)} \quad (2.3)$$

5) คำนวณค่าโลคอลเอาท์ไลเออร์แฟคเตอร์โดยการเปรียบเทียบความหนาแน่นของจุดที่เราสนใจเปรียบเทียบกับจำนวนเพื่อนบ้านใกล้เคียง ดังสมการที่ 2.4

$$LOF_k(p) = \frac{\sum_{o \in N_k(p)} \frac{lrd_k(q)}{lrd_k(p)}}{|N_k(p)|} \quad (2.4)$$

โดยที่ $d(p,q)$ คือ ระยะห่างระหว่างจุด p ถึงจุด q
$dist_k(p)$ คือ ระยะทางจากจุด p ถึงจุด k
$N_k(p)$ คือ จุดข้อมูลที่ระยะทางจุด p น้อยกว่า $dist_k(p)$
$lrd_k(p)$ คือ ส่วนกลับของค่าเฉลี่ยของระยะทางที่เข้าถึงได้ของจุดข้อมูล p และ k เพื่อนบ้านที่ใกล้ที่สุด
$dist_k(q)$ คือ ระยะทางจากจุด q ถึงจุด k
$lrd_k(q)$ คือ ส่วนกลับของค่าเฉลี่ยของระยะทางที่เข้าถึงได้ของจุดข้อมูล q และ k เพื่อนบ้านที่ใกล้ที่สุด

ค่าโลคอลเอาท์ไลเออร์แฟคเตอร์ เป็นอัตราส่วนของผลรวมความหนาแน่นของเพื่อนบ้านเทียบกับจำนวนเพื่อนบ้านใกล้เคียง กรณีที่ผลรวมความหนาแน่นของเพื่อนบ้านมีค่าน้อยกว่าจำนวนเพื่อนบ้านใกล้เคียงส่งผลให้ค่าโลคอลเอาท์ไลเออร์แฟคเตอร์มีค่ามากกว่า -1.5 เป็นค่าที่ปรกติ กรณีที่ผลรวมความหนาแน่นของเพื่อนบ้านมีค่ามากกว่าจำนวนเพื่อนบ้านใกล้เคียง ส่งผลให้ค่าโลคอลเอาท์ไลเออร์แฟคเตอร์มีค่าน้อยกว่า -1.5 เป็นค่าที่ผิดปรกติ จะเห็นได้ว่าวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์จะขึ้นอยู่กับเพื่อนบ้านใกล้เคียง (Breunig et al., 2000)



รูปที่ 2.4 ระยะทางของเพื่อนบ้านใกล้เคียง K ตัว ของวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ (ที่มา: Breunig et al., 2000)

2.3.2 วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง (One-Class Support Vector Machine: OC-SVM)

วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่งเป็นเทคนิคหนึ่งของวิธีซัพพอร์ตเวกเตอร์แมชชีน Amer et al. (2013) ได้กล่าวว่าการทำงานของวิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่งต่างจาก วิธีซัพพอร์ตเวกเตอร์แมชชีนแบบเดิม คือวิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง พยายามสร้างขอบเขตการตัดสินใจที่ทำให้เกิดการแยกข้อมูลส่วนใหญ่จากจุดกำเนิดมีเพียงจุดเล็ก ๆ ของข้อมูลที่อยู่อีกด้านของขอบเขตการตัดสินใจ จุดข้อมูลเหล่านั้นถือเป็นค่าที่ผิดปรกติใช้ฟังก์ชันการแปลงที่กำหนดโดยเคอร์เนลเพื่อทำให้ข้อมูลมีพื้นที่มิติสูงขึ้น ถูกกำหนดดังสมการที่ 2.5

$$g(x) = w^T \phi(x) - \rho \tag{2.5}$$

โดยที่ $\phi(x)$ คือ ฟังก์ชันการแปลงที่กำหนดโดยเคอร์เนล w คือเวกเตอร์ที่ตั้งฉากกับขอบเขตการตัดสินใจ และ ρ คือระยะเอนเอียง จากนั้นสมการที่ 2.6 จะแสดงฟังก์ชันการตัดสินใจที่วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง ใช้เพื่อระบุจุดที่ปรกติ โดยฟังก์ชันจะให้ค่าบวกสำหรับค่าที่ปรกติ และค่าที่เป็นลบสำหรับค่าที่ผิดปรกติ

$$f(x) = \text{sgn}(g(x)) \tag{2.6}$$

โดยขั้นตอนวิธีจะเรียนรู้ขอบเขตการตัดสินใจไฮเปอร์เพลน (Hyperplane) ที่แยกข้อมูลส่วนใหญ่ออกจากจุดกำเนิดของวิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง ผลลัพธ์ของขั้นตอนจะได้เป็นป้ายกำกับแบบไบนารีที่ระบุว่าจุดนั้นเป็นปรกติหรือไม่ สมการที่ 2.7 แสดงเป้าหมายหลักของวิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง

$$min_{w,\xi,\rho} \frac{1}{2}\|w\|^2 - \rho + \frac{1}{\upsilon n}\sum_{i=0}^{n}\xi_i \tag{2.7}$$

$$s.t. \begin{cases} (w^T, \phi(x_i)) \geq \rho - \xi_i \\ \xi_i \geq 0 \end{cases}$$

โดยที่ ξ_i คือเป็นตัวแปรหย่อนสำหรับจุดที่ i ที่อนุญาตให้อยู่อีกด้านหนึ่งของขอบเขตการตัดสินใจ n คือขนาดของชุดข้อมูล และ υ คือพารามิเตอร์ที่ทำให้เป็นมาตรฐาน จากทฤษฎีทางคณิตศาสตร์ สามารถระบุได้โดยระยะทางจุดกำเนิดถึงของเขตการตัดสินใจ โดยกำหนดดังสมการที่ 2.8

$$g(x) = 0 \tag{2.8}$$

ในคำอธิบายนี้ระยะทางของข้อมูลใด ๆ สามารถคำนวณขอบเขตการตัดสินใจดังสมการที่ 2.9

$$d(x) = \frac{|g(x)|}{\|w\|} \tag{2.9}$$

ดังนั้นระยะทางที่ขั้นตอนวิธีนี้พยายามจะทำให้ได้สูงสุด สามารถทำได้โดยการนำ ρ แทนในจุดกำเนิด เพื่อกำหนดความกว้างของเส้นขอบในสมการ $\frac{\rho}{\|w\|}$ นอกจากนี้ยังสามารถลดขนาดความกว้าง ของสมการได้โดย $\frac{\|w\|^2}{2} - \rho$

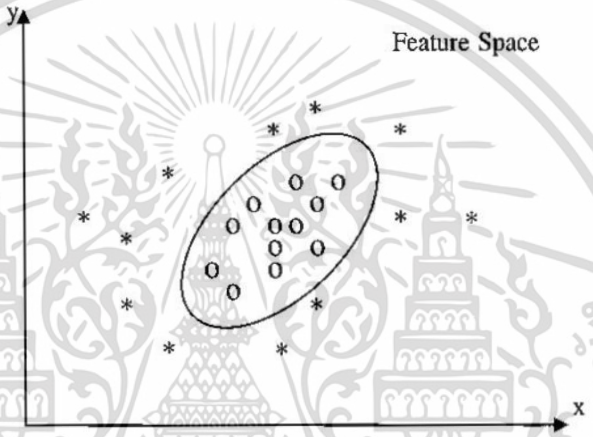
ส่วนที่สองของวัตถุประสงค์หลักคือการย่อขนาดตัวแปรหย่อนให้น้อยที่สุดสำหรับทุกจุด และ υ เป็นพารามิเตอร์การทำให้เป็นมาตรฐานหมายถึงขอบเขตบนของค่าที่ผิดปรกติและขอบเขตล่างของค่าที่ปรกติของเส้นซัพพอร์ตเวกเตอร์การเปลี่ยนแปลง υ ควบคุมการ trade-off ระหว่าง ξ และ ρ ด้วยเหตุนี้ วัตถุประสงค์หลักจะถูกแปลงเป็นสมการที่ 2.10 การแปลงทำให้วิธีซัพพอร์ตเวกเตอร์แมชชีน สามารถใช้เคอร์เนล ได้เช่นเดียวกับการลดจำนวนของตัวแปรให้เป็นหนึ่งเวกเตอร์ โดยทั่วไปแล้ว Quadratic Programming (QP) จะเพิ่มประสิทธิภาพ

$$min_{\alpha} = \frac{\alpha^T Q \alpha}{2} \tag{2.10}$$

$$s.t.: 0 \le \alpha_i \le \frac{1}{\upsilon n}, \sum_{i=1}^{n} \alpha_i = 1$$

โดยที่ Q คือเคอร์เนลเมทริกซ์ และ α เป็นตัวคูณลากรานจ์

สำหรับเคอร์เนล (Kernel) ในความเป็นจริงนั้นข้อมูล 2 กลุ่ม ไม่ได้วางตัวในพื้นที่ข้อมูลคุณลักษณะ และไม่สามารถแบ่งได้โดยเส้นตรง แต่ข้อมูลอาจจะจับกลุ่มกันในตำแหน่งต่าง ๆ ดังนั้นจึงเป็นปัญหาทำให้ไม่สามารถที่จะใช้สมการซัพพอร์ตเวกเตอร์แมชชีนแบบเชิงเส้นได้ ดังนั้นจะต้องมีเครื่องมือมาช่วยให้ข้อมูลเหล่านั้นเรียงตัวใหม่ในพื้นที่เรียกว่า พื้นที่หลายมิติ (Higher Dimensional Space)



รูปที่ 2.5 รูปแบบการวางตัวที่ไม่สามารถแบ่งด้วยเส้นตรงได้
(ที่มา: สุรวัชร และ สายชล, 2560)

ในขั้นตอนวิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง ใช้ชุดฟังก์ชันทางคณิตศาสตร์ที่กำหนดเป็นเคอร์เนล โดยฟังก์ชันของเคอร์เนล คือ การรับข้อมูลเป็นอินพุต (Input) และแปลงเป็นรูปแบบที่ต้องการ มีรูปแบบที่แตกต่างกันดังนี้

1. Linear kernel คำนวณได้ดังสมการที่ 2.11

$$k(x_i, x_j) = x_i^T x_j \tag{2.11}$$

2. Polynomial kernel เป็นที่นิยมในการประมวลลภาพ (Image Processing) คำนวณได้ดังสมการที่ 2.12

$$k(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0 \tag{2.12}$$

3. Radial basis function: RBF เป็นเคอร์เนลที่นิยมใช้เนื่องจากความสามารถในการจับความสัมพันธ์ที่ไม่ใช่เชิงเส้นในข้อมูล คำนวณได้ดังสมการที่ 2.13

$$k(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right), \gamma > 0 \quad (2.13)$$

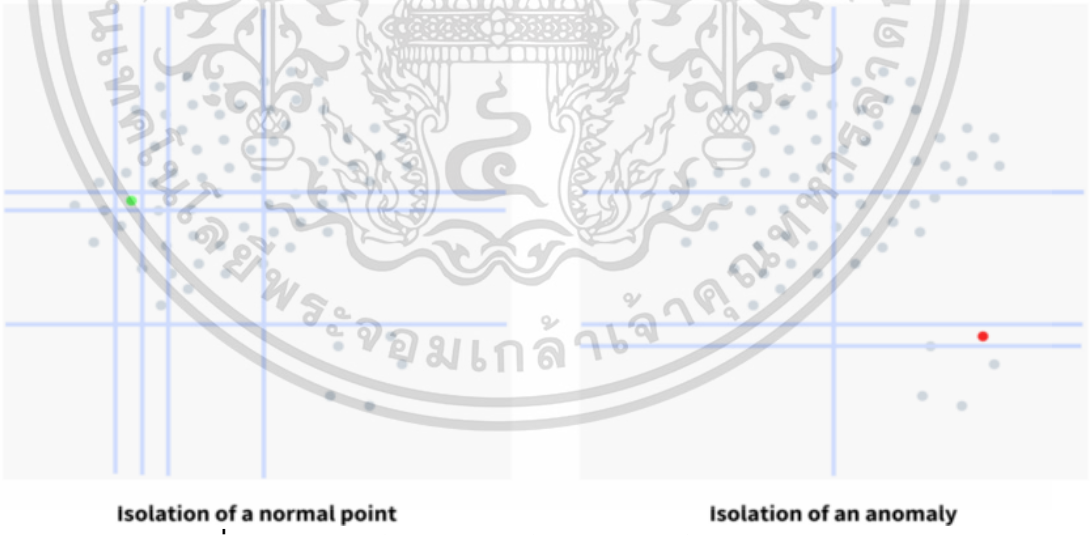
4. Sigmoid kernel ใช้เป็นพร็อกซี (proxy) สำหรับโครงข่ายประสาทเทียม คำนวณได้ดังสมการที่ 2.14

$$k\left(x_i - x_j\right) = exp(\gamma x_i^T x_j + r) \quad (2.14)$$

โดยที่ γ, r และ d เป็นค่าพารามิเตอร์ของเคอร์เนล (Hsu et al., 2016)

2.3.3 วิธีไอโซเลชั่นฟอเรส (Isolation Forest: IF)

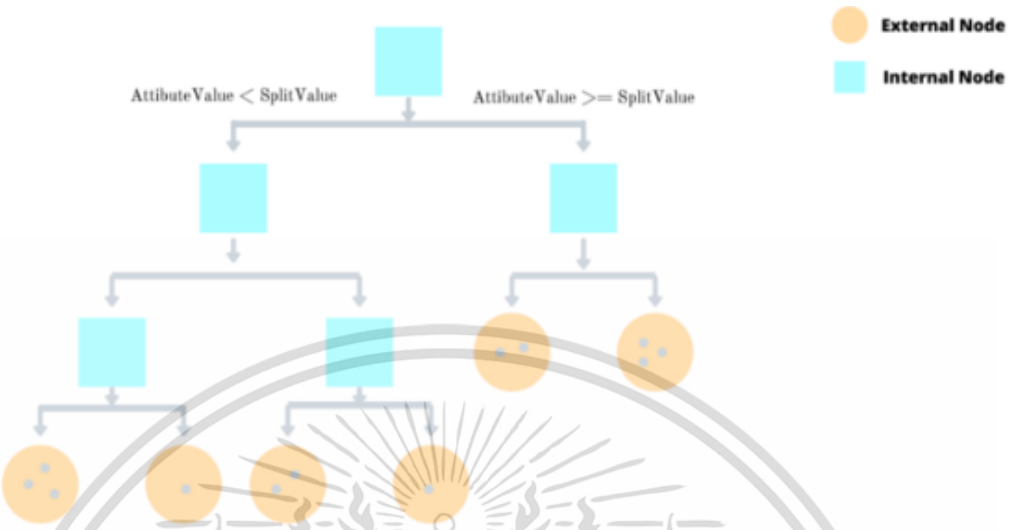
วิธีการตรวจจับสิ่งผิดปรกติที่มีรากฐานมาจากวิธีต้นไม้ตัดสินใจ (Decision Tree) โดยเริ่มต้นจากการสุ่มคุณลักษณะ (Attribute) และแบ่งข้อมูล (Partition) ระหว่างค่าต่ำสุด และค่าสูงสุดเพื่อแยกตัวอย่าง โดยจะแบ่งข้อมูลไปเรื่อย ๆ จนกระทั่งข้อมูลแต่ละตัวจะแยกจากกันโดยสมบูรณ์ วิธีไอโซเลชั่นฟอเรสถูกสร้างขึ้นจากการเพิ่มขึ้นของจำนวนไอโซเลชั่นทรี (Isolation Tree) ที่ถูกแยกด้วยคุณลักษณะต่าง ๆ ที่แตกต่างกัน (Farzad and Gulliver, 2020)



รูปที่ 2.6 การแบ่งข้อมูลปรกติ (ซ้าย) การแบ่งข้อมูลผิดปรกติ (ขวา)
(ที่มา: อาณัติชัย, 2565)

เมื่อข้อมูลจุดสีเทากระจายตัวในลักษณะดังรูปที่ 2.6 หากต้องการจะแบ่งข้อมูลออกจากกันสามารถทำได้โดยสร้างเส้นแบ่งขึ้นมาอาจจะเป็นแนวตั้งหรือแนวนอนก็ได้ แต่ต้องแบ่งจนกว่าจุดที่สนใจจะถูกแยกออกจากจุดอื่นโดยสิ้นเชิง ซึ่งพิจารณารูปทางซ้ายจะเห็นได้ว่าข้อมูลปรกติจะอยู่กระจุกตัวกับข้อมูลจุดอื่น ๆ ทำให้ต้องใช้จำนวนเส้นในการแบ่งค่อนข้างมาก เปรียบเทียบกับข้อมูล

ผิดปกติรูปทางขวาที่จะอยู่แยกจากข้อมูลส่วนใหญ่ ทำให้ใช้จำนวนเส้นแบ่งที่น้อยกว่าก็สามารถแยกออกมาได้



รูปที่ 2.7 ต้นไม้ตัดสินใจ (Decision Tree)
(ที่มา: อาณัติชัย, 2565)

จากรูปที่ 2.7 จะเห็นได้ว่าเส้นแบ่งแต่ละเส้นก็คือเส้นที่ตัดแบ่งข้อมูลทั้งหมดออกเป็นกิ่งซ้าย และกิ่งขวาแล้วทำการแบ่งไปจนกระทั่งถึงจุดที่ทุกอย่างแยกออกจากกัน ทำให้ได้ต้นไม้ที่มีกิ่งแยกข้อมูลออกจากกัน หลังจากนั้นจะพบว่าข้อมูลปรกติจะใช้จำนวนชั้นของต้นไม้ที่ลึกมากในการแบ่งข้อมูลให้เป็นอิสระจากกัน แต่ข้อมูลผิดปรกติจะถูกกรองตั้งแต่ชั้นแรก ๆ ของต้นไม้ทำให้ความลึกของข้อมูลที่ผิดปรกติจะตื้นกว่าข้อมูลอื่น ๆ

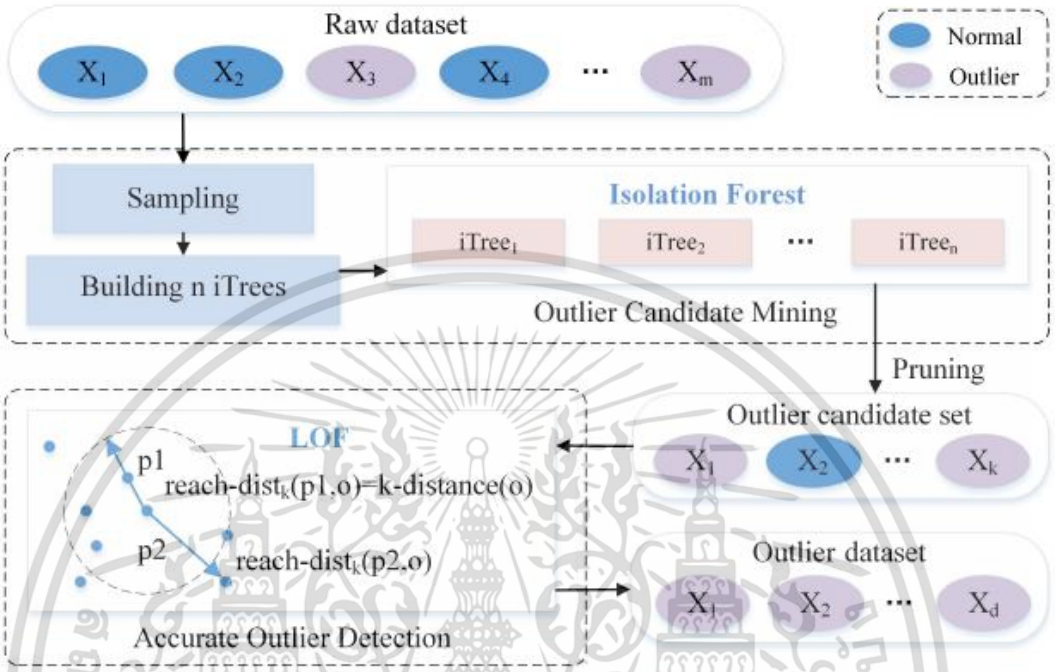
เมื่อทราบความลึกของต้นไม้แล้วทำให้สามารถคำนวณเป็นคะแนนความผิดปรกติ (Anomaly score) เพื่อใช้ในการแยกประเภทของข้อมูลได้ ซึ่งคะแนนความความผิดปรกตินั้นจะมีค่าน้อยกว่า -0.5 จะหมายถึงข้อมูลที่มีแนวโน้มผิดปรกติ ส่วนข้อมูลที่มีค่ามากกว่า -0.5 จะถือว่าเป็นข้อมูลทั่วไปที่ไม่มีความผิดปรกติ (Hui, 2021)

2.3.4 วิธีไอเอฟ-แอลโอเอฟ (Isolation Forest- Local Outlier Factor: IF-LOF)

เป็นวิธีแก้ไขปัญหาการตรวจจับสิ่งผิดปรกติที่อ่อนไหวต่อความผิดปรกติที่มีความหลากหลาย (Global Outlier) และใช้เวลาในการประมวลผลนาน โดยจะทำการตัด (Prune) ชุดข้อมูลแทนการใช้ชุดข้อมูลดั้งเดิม และนำมาใช้เป็นแหล่งข้อมูล (Data Source) จึงสามารถลดปริมาณข้อมูลที่ต้องประมวลผลได้เป็นอย่างมาก (Zhangyu et al., 2019) โดยมีวิธีการดังนี้

1) นำข้อมูลดิบไปใช้ในวิธีไอโซเลชั่นฟอเรสซึ่งถูกสร้างขึ้นจากการเพิ่มขึ้นของจำนวนไอโซเลชั่นทรี (Isolation Tree) ที่ถูกแยกด้วยคุณลักษณะต่าง ๆ ที่แตกต่างกัน จากนั้นคำนวณคะแนนความผิดปรกติจากความลึกของต้นไม้
2) ทำการตัดกิ่ง (Pruning) โดยตัดจุดข้อมูลปรกติบางส่วนออกตามเกณฑ์การตัดแต่งกิ่งเพื่อให้ได้ชุดข้อมูลผิดปรกติที่เหลืออยู่

3) คำนวณค่าโลคอลเอาท์ไลเออร์แฟคเตอร์ แต่ละจุดข้อมูลที่มาจากชุดข้อมูลผิดปรกติที่เหลืออยู่ และเลือก d จุดแรกที่มีค่าโลคอลเอาท์ไลเออร์แฟคเตอร์สูงเป็นสิ่งผิดปรกติ



รูปที่ 2.8 แผนภาพการทำงานของวิธี ไอเอฟ-แอลโอเอฟ
(ที่มา: Zhangyu et al., 2019)

2.4 การวิเคราะห์จัดกลุ่ม (Cluster Analysis)

การวิเคราะห์จัดกลุ่ม (Cluster Analysis) เป็นเทคนิคการแบ่งกลุ่มหน่วยข้อมูล หรือเป็นการแบ่งคน สัตว์ สิ่งของ องค์กร ฯลฯ ออกเป็นกลุ่มย่อยอย่างน้อย 2 กลุ่ม โดยมีหลักเกณฑ์ในการแบ่งดังนี้ “ให้หน่วยที่อยู่ในกลุ่มเดียวกันมีลักษณะที่สนใจเหมือนกันหรือคล้ายกัน แต่หน่วยที่อยู่ต่างกลุ่มกันจะมีลักษณะที่สนใจต่างกัน” (กัลยา, 2552)

การจัดกลุ่มข้อมูลเป็นวิธีการวิเคราะห์ข้อมูลซึ่งอาศัยการเรียนรู้ของเครื่อง โดยจะแบ่งชุดข้อมูล (มักจะเป็นเวกเตอร์) ออกเป็นกลุ่ม (Cluster) นำข้อมูลที่มีคุณลักษณะเหมือนกันหรือคล้ายกันจัดใส่ในกลุ่มเดียวกัน ขั้นตอนวิธีที่ใช้ในการแบ่งกลุ่มจะอาศัยความเหมือน (Similarity) หรือความใกล้ชิด (Proximity) โดยคำนวณจากการวัดระยะระหว่างเวกเตอร์ของข้อมูลเข้า โดยวิธีการคำนวณระยะทางแบบต่าง ๆ

2.4.1 การคำนวณค่าระยะห่างในข้อมูล

1. Euclidean Distance

Euclidean Distance เป็นการวัดระยะห่างระหว่างจุดสองจุด เช่นจุด p และจุด q ใน Cartesian Coordinates ถ้า $p=(p_1,p_2,...,p_n)$ และ $q=(q_1,q_2,...,q_n)$ เป็นจุดสอง

จุดบนปริภูมิยุคลิด n มิติ ระยะทางระหว่างจุด p กับ q หรือ q กับ p คำนวณได้จากสมการที่ 2.15 (Deza and Deza, 2009)

$$d(p,q)=d(p,q)=\sqrt{(q_1-p_1)^2+(q_2-p_2)^2+...+(q_n-p_n)^2}=\sqrt{\sum(q_i-p_i)^2} \quad (2.15)$$

โดยที่ $d(p,q)$ คือ ระยะทางจาก p ไปจนถึง q
$d(p,q)$ คือ ระยะทางจาก q ไปจนถึง p
n คือ จำนวนมิติข้อมูล

2. Manhattan Distance

Manhattan Distance หรือเรียกอีกอย่างว่า Taxicab โดยวิธีการแบบ Manhattan วัดระยะระหว่างจุดสองจุดโดยใช้ผลรวมของค่าสัมบูรณ์ของผลต่างของแต่ละมิติ โดยรูปแบบทั่วไปสามารถอธิบายได้ดังนี้

กำหนดระยะทาง d เป็นระยะทางระหว่างเวกเตอร์ p และ q ที่อยู่ใน n มิติ ระยะทางระหว่าง q และ p ซึ่ง $p=(p_1,p_2,...,p_n)$ และ $q=(q_1,q_2,...,q_n)$ อธิบายได้ด้วยสมการที่ 2.16 ดังต่อไปนี้ (Krause, 1987)

$$d(p,q)=\|p-q\|=\sum_{i=1}^{n}|p_i-q_i| \quad (2.16)$$

3. Minkowski Distance

Minkowski Distance เป็นรูปแบบทั่วไปของวิธี Euclidean Distance และ Manhattan Distance นิยามของวิธี Minkowski มีดังนี้กำหนดระยะทาง d เป็นระยะทางระหว่างเวกเตอร์ p และ q ที่อยู่ใน n มิติ ระยะทาง $p=(p_1,p_2,...,p_n)$ และ $q=(q_1,q_2,...,q_n)$ คำนวณได้จากสมการที่ 2.17 (Myatt and Johnson, 2009)

$$d(p,q)=\sqrt{\sum_{i=1}^{n}|p_i-q_i|^{\lambda}} \quad (2.17)$$

โดยที่ถ้าค่า λ มีค่าเป็น 1 จะได้ผลการคำนวณเหมือนกับ Manhattan Distance ถ้าเป็น 2 จะได้ผลเหมือนกับ Euclidean Distance

2.4.2 เทคนิคการจัดกลุ่ม

เทคนิคที่ใช้นิยมใช้กันมากมี 2 เทคนิค คือ

1. วิธี K-Means Clustering

K-Means เป็นหนึ่งในอัลกอริทึมเทคนิคการเรียนรู้โดยไม่มีผู้สอนที่มีความง่ายที่สุด และเป็นการแก้ปัญหาการจัดกลุ่มที่รู้จักกันโดยทั่วไปโดยอัลกอริทึม K-Means มีหลักการคือ แบ่งกลุ่มวัตถุออกเป็นกลุ่ม K (Partition) แทนแต่ละกลุ่มด้วยค่าเฉลี่ยของกลุ่ม และในการวัดระยะห่างของข้อมูลในกลุ่มเดียวกันจะใช้ค่าเฉลี่ยนี้เป็นจุดศูนย์กลาง (Centroid) ของกลุ่ม โดยมีวิธีการทำงาน (Hartigan and Worg, 1979) ดังนี้

1) คำนวณหาค่าเฉลี่ยโดยเริ่มจากการกำหนดจำนวนกลุ่ม (K) ที่ต้องการและกำหนดจุดศูนย์กลางเริ่มต้นจำนวน K จุด

2) ในการกำหนดจุดศูนย์กลางเริ่มต้นของแต่ละกลุ่มนี้ ควรจะถูกกำหนดด้วยวิธีที่เหมาะสม เพราะตำแหน่งจุดศูนย์กลางเริ่มต้นที่แตกต่างกัน ทำให้ได้ผลลัพธ์สุดท้ายแตกต่างกันตามไปด้วย ควรจะกำหนดจุดศูนย์กลางนี้ให้ห่างจากจุดศูนย์กลางอื่น ๆ

3) จากนั้นสร้างกลุ่มข้อมูลและความสัมพันธ์กับจุดศูนย์กลางที่ใกล้มากที่สุด โดยแต่ละจุดจะถูกกำหนดไปยังจุดศูนย์กลางที่ใกล้เคียงที่สุดจนครบหมดทุกจุด และคำนวณจุดศูนย์กลางใหม่ โดยการหาค่าเฉลี่ยของทุกวัตถุที่อยู่ในกลุ่ม หากจุดศูนย์กลางในแต่ละกลุ่มถูกเปลี่ยนตำแหน่ง จะได้จุดที่มีความสัมพันธ์กับกลุ่มใหม่ และใกล้กับจุดศูนย์กลางใหม่ ทำซ้ำแบบนี้ไปเรื่อย ๆ จะพบว่าผลลัพธ์จากการทำซ้ำแบบนี้ ทำให้จุดศูนย์กลางเปลี่ยนตำแหน่งทุกรอบจนกระทั่งเป็นจุดศูนย์กลางจำนวน K จุดที่ไม่มีการเปลี่ยนแปลงจึงจะสิ้นสุดกระบวนการ

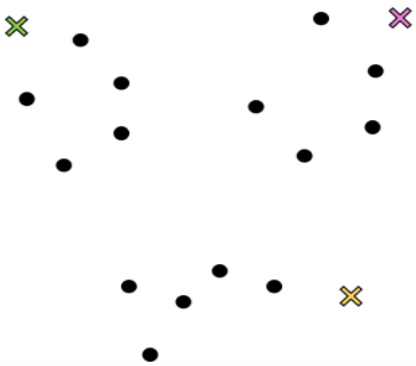
ข้อเด่นของอัลกอริทึมจัดกลุ่มแบบ K-Means

1) ง่ายและนิยมนำไปใช้งาน
2) จัดกลุ่มข้อมูลได้รวดเร็ว

ข้อด้อยของอัลกอริทึมจัดกลุ่มแบบ K-Means

1) การกำหนดจำนวนกลุ่ม และตัวแทนกลุ่มเริ่มต้น มีผลต่อประสิทธิภาพการจัดกลุ่มทั้งในเชิงเวลา และความถูกต้อง
2) ขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบ K-Means จะมีปัญหากับข้อมูลที่มี Outliers
3) ข้อมูลมีโอกาสเป็นสมาชิกเพียงกลุ่มใดกลุ่มหนึ่งเท่านั้น
4) ผลลัพธ์ของการจัดกลุ่มที่ได้เป็นแบบ Local Optimal

ตัวอย่างการจัดกลุ่มโดยอัลกอริทึมจัดกลุ่มแบบ K-Means เมื่อแบ่งข้อมูลเป็น 3 กลุ่ม แสดงดังรูปที่ 2.9

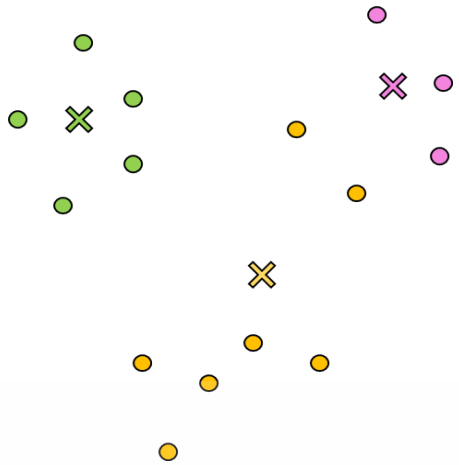


รูปที่ 2.9 การกำหนดจุดข้อมูลเพื่อใช้ในการจัดกลุ่มข้อมูลแบบ K-Means (Kumar, 2002)

จากรูปที่ 2.9 สัญลักษณ์วงกลมสีดำคือข้อมูลใด ๆ สัญลักษณ์กากบาท สีเขียว สีชมพู และสีเหลือง คือจุดข้อมูลที่สุ่มได้ และแทนจุดศูนย์กลางของกลุ่มข้อมูล ซึ่งในกรณีนี้กำหนดจำนวนกลุ่มเป็น 3 กลุ่ม

รูปที่ 2.10 ตัวอย่างการจัดกลุ่มข้อมูล โดยพิจารณาค่าระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางของกลุ่มข้อมูลที่สุ่มได้ในครั้งแรก (Kumar, 2002)

จากรูปที่ 2.10 สัญลักษณ์วงกลมสีเขียว สีชมพู และสีเหลือง คือข้อมูลใด ๆ ที่จะถูกจัดกลุ่ม โดยพิจารณาค่าระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางของกลุ่มข้อมูลที่สุ่มได้ในครั้งแรก สัญลักษณ์กากบาทสีเขียว สีชมพู และสีเหลือง คือ จุดศูนย์กลางของกลุ่มข้อมูลที่สุ่มได้ในครั้งแรก



รูปที่ 2.11 ตัวอย่างการปรับจุดศูนย์กลางของข้อมูลแต่ละกลุ่มไปยังตำแหน่งที่ครอบคลุมข้อมูลบริเวณใกล้เคียงภายในกลุ่มเดียวกันให้มากที่สุด โดยการหาค่าเฉลี่ยของข้อมูลภายในกลุ่มเดียวกันให้ได้มากที่สุด (Kumar, 2002)

จากรูปที่ 2.11 สัญลักษณ์วงกลมสีเขียว สีชมพู และสีเหลือง คือข้อมูลใด ๆ สัญลักษณ์กากบาท สีเขียว สีชมพู และสีเหลือง คือ จุดศูนย์กลางของกลุ่มข้อมูลที่ได้รับการปรับให้อยู่ในตำแหน่งที่ครอบคลุมข้อมูลในกลุ่มเดียวกันมากขึ้น

รูปที่ 2.12 การจัดกลุ่มที่เสร็จสมบูรณ์ โดยจุดศูนย์กลางที่ได้จะอยู่ในตำแหน่งที่เป็นตัวแทนของข้อมูลแต่ละกลุ่ม (Kumar, 2002)

จากรูปที่ 2.12 สัญลักษณ์วงกลมสีเขียว สีชมพู และสีเหลือง คือข้อมูลใด ๆ สัญลักษณ์กากบาท สีเขียว สีชมพู และสีเหลือง คือ จุดศูนย์กลางของกลุ่มข้อมูลที่เป็นตัวแทนที่เหมาะสมที่สุดของข้อมูลในแต่ละกลุ่ม

2. วิธี Hierarchical Cluster Analysis

Johnson (1967) เป็นผู้นำเสนอวิธี Hierarchical Cluster ซึ่งเป็นหนึ่งในอัลกอริทึมที่นิยมใช้กันมากในการจัดกลุ่ม หรือ จัดกลุ่มตัวแปร การแบ่งกลุ่มประเภทนี้เป็นวิธีการจัดกลุ่มแบบขั้นตอนโครงสร้างกลุ่มเหมือนต้นไม้เรียกว่า “เดนโดแกรม (Dendrogram)” สร้างโดยใช้

วิธีการจัดกลุ่มแบบรวมกัน (Agglomerative Clustering Method) หรือ สร้างโดยใช้วิธีการจัดกลุ่มแบบแยกกัน (Divisive Clustering Method) ของกลุ่มต่างๆ ที่ปรากฏ

2.1 วิธีการจัดกลุ่มแบบรวมกัน (Agglomerative Clustering Method) เริ่มต้นด้วยค่าสังเกตแต่ละค่าถือเป็น 1 กลุ่ม และกลุ่มที่อยู่ใกล้กันมากที่สุด 2 กลุ่ม จะถูกนำมารวมกันเป็นกลุ่มใหม่เพียงกลุ่มเดียว ดังนั้นจำนวนกลุ่มในชุดข้อมูลจะลดลง ในแต่ละขั้นตอน จนในที่สุดระเบียน (Record) ทั้งหมดจะถูกนำมารวมกันเป็นกลุ่มขนาดใหญ่เพียงกลุ่มเดียว

2.2 วิธีการจัดกลุ่มแบบแยกกัน (Divisive Clustering Method) เริ่มต้นด้วยมีระเบียนทั้งหมดในกลุ่มขนาดใหญ่เพียงกลุ่มเดียว หลังจากนั้นระเบียนที่เหมือนกันจะเริ่มแยกออกจากกันไปอยู่ในกลุ่มแตกต่างกัน จนกระทั่งแต่ละ ระเบียนแสดงการเป็นสมาชิกกลุ่มของตัวเอง เนื่องจากโปรแกรมคอมพิวเตอร์ส่วนใหญ่ใช้วิธีการจัดกลุ่มแบบรวมกัน ดังนั้นจึงเน้นวิธีการจัดกลุ่มแบบรวมกัน

2.4.3 การกำหนดจำนวน K ที่เหมาะสมในการทำ K-Means Clustering

1. Elbow Method

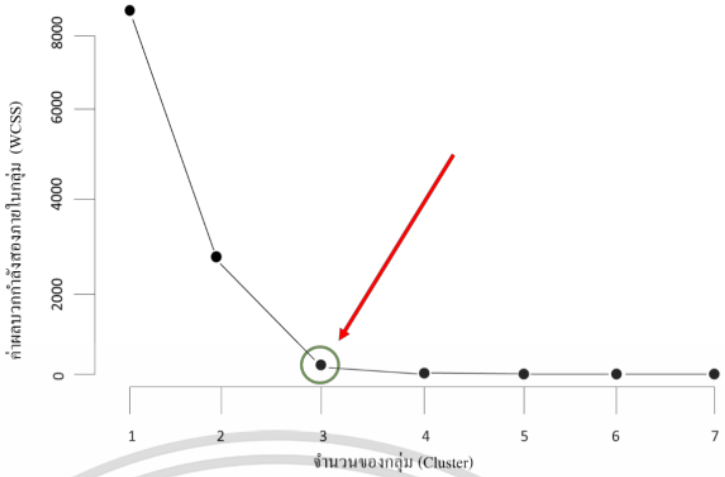
Elbow Method คือวิธีการวัดค่าความคลาดเคลื่อนของผลรวมของระยะห่างระห่างวัตถุ (Object) กับจุดศูนย์กลางเริ่มต้น (Centroid) เรียกว่าผลบวกกำลังสอง (Sum Of Square) หรือ เรียกอีกชื่อได้ว่า ค่าผลบวกกำลังสองภายในกลุ่ม (With In Cluster of Square: WCSS) อธิบายได้ด้วยสมการที่ 2.18 ดังต่อไปนี้ (ศศิวุฒิ, 2563)

$$WCSS = \sum_{C_k}^{C_n} \left(\sum_{d_i in C_i}^{d_m} DISTANCE(d_i, C_k)^2 \right) \qquad (2.18)$$

โดยที่ C_k คือ จุดศูนย์กลาง (Centroid) ของกลุ่ม (Cluster)

d_i คือ จุดข้อมูลในแต่ละกลุ่ม (Cluster)

ในการจัดกลุ่มแต่ละรอบค่าที่ทำการคำนวณจะมีค่าลดลงเรื่อย ๆ จากจำนวนกลุ่มที่เยอะขึ้นเพราะว่าสมาชิกในแต่ละกลุ่มจะลดลงไปเรื่อย ๆ ดังนั้นค่าความแปรปรวนภายในกลุ่ม (SSE) จะทำให้เกิดความโค้งที่เรียบขึ้นเรื่อย ๆ จุดที่เหมาะสมของจำนวนกลุ่มคือจุดที่กราฟมีลักษณะ “หักศอก” ที่สุด ดังนั้นจากการคำนวณหาจำนวนกลุ่มที่เหมาะสมที่สุดจึง สามารถหาได้จากระยะทางที่ไกลที่สุดนับจากเส้นตรงระหว่าง สอง จุดกับเส้นโค้ง จุดหักศอกที่เกิดใน รูปที่ 2.13 มีการเปลี่ยนแปลงของจุดที่เห็นชัดที่สุดคือจุดที่ 3 หมายความว่าควรจะแบ่งจำนวนของกลุ่มอยู่ที่จำนวน 3 กลุ่ม



รูปที่ 2.13 จุดที่เหมาะสมของจำนวน Clusters
ที่มา (ศศิวุฒิ, 2563)

2. Silhouette Coefficient

Rousseeuw (1987) ได้กล่าว และนำเสนอไว้ว่า Silhouette เป็นเครื่องมือที่ใช้สำหรับการจัดกลุ่มข้อมูลใช้เพื่อเลือกจำนวนกลุ่มที่เหมาะสมโดยหาค่า Silhouette จะใช้เปรียบเทียบว่าการจัดกลุ่มแบบใดมีประสิทธิภาพการจัดกลุ่มได้ดีที่สุด โดยการวัดประสิทธิภาพของข้อมูลที่ได้ทำการจัดกลุ่มซึ่งจะพิจารณาจากค่า Silhouette Coefficient โดยใช้ค่าเฉลี่ยของระยะห่างระหว่างจุด Si กับจุดต่าง ๆ ภายในกลุ่มเดียวกัน ส่วนด้วยระยะห่างน้อยที่สุดของจุด Si กับจุดต่าง ๆ ในแต่ละกลุ่มจะได้ดังสมการที่ 2.19

$$S(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{2.19}$$

โดยที่ $a(i)$ คือ ระยะห่างเฉลี่ยของข้อมูลที่ i กับข้อมูลอื่นๆทั้งหมดภายในกลุ่มเดียวกัน

$b(i)$ คือ ระยะห่างเฉลี่ยของข้อมูลที่ i กับข้อมูลทั้งหมด

และค่า S_i ที่ได้จะอยู่ในช่วง $-1 \le S(i) \le 1$ ซึ่งค่าของ $S(i)$ มีค่าใกล้ 1 แสดงว่าข้อมูล ที่ i ถูกจัดให้อยู่ในกลุ่มมีความเหมาะสมแล้ว ถ้า $S(i)$ มีค่าใกล้ -1 แสดงว่าข้อมูลที่ i ไม่เหมาะสมกับกลุ่มที่จัดไว้ ควรจัดให้อยู่ในกลุ่มอื่น ๆ และถ้า $S(i)$ มีค่าใกล้ 0 แสดงว่าข้อมูลที่ i ยังไม่เหมาะสมกับกลุ่มที่จัดไว้

2.5 แนวคิดและทฤษฎีที่เกี่ยวข้องกับการวิเคราะห์ทางสถิติ (Statistics)

2.5.1. สถิติเชิงพรรณนา (Descriptive Statistics)

สถิติที่ใช้ในการวิเคราะห์ข้อมูลพื้นฐานในงานวิจัยเชิงปริมาณ เริ่มที่การวิเคราะห์เพื่อบรรยายลักษณะข้อมูลทั่วไป และวิเคราะห์เพื่อตอบคำถามหรือวัตถุประสงค์การวิจัยในกรณีที่เป็นคำถามเชิงพรรณนา (สุทิน, 2560) โดยการวิเคราะห์เชิงพรรณนาที่ใช้ประกอบด้วย

1. ค่าร้อยละหรือเปอร์เซ็นต์ (Percentage)

ค่าร้อยละหรือเปอร์เซ็นต์ คือ การคำนวณเพื่อหาสัดส่วนของข้อมูลในแต่ละตัวเทียบกับข้อมูลทั้งหมดโดยให้ข้อมูลรวมทั้งหมดมีค่าเท่ากับ 100 ใช้สัญลักษณ์ % แทนที่คำว่าร้อยละ ซึ่งใช้วิเคราะห์ตัวแปรประเภทมาตรานามบัญญัติ (Nominal Scale) หรือมาตราอันดับ (Ordinal Scale) ซึ่งมีสูตรคำนวณดังสมการที่ 2.20

$$\text{ร้อยละ} = \frac{x}{n} \times 100 \tag{2.20}$$

เมื่อ x คือ จำนวนข้อมูลหรือความถี่
n คือ ขนาดตัวอย่าง

2. ค่าเฉลี่ย (Mean)

ค่าเฉลี่ย คือ ค่ากลางของข้อมูลรูปแบบหนึ่งจะใช้ค่าเฉลี่ยเป็นตัวแทนของข้อมูลที่นำมาคำนวณ ใช้วิเคราะห์ตัวแปรประเภทมาตราช่วง (Interval Scale) หรือมาตราส่วนอัตราส่วน (Ratio Scale) ซึ่งมีสูตรคำนวณดังสมการที่ 2.21

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.21}$$

เมื่อ $\bar{x}$ คือ ค่าเฉลี่ยของข้อมูล
x_i คือ คะแนนของตัวอย่างชุดที่ i
n คือ ขนาดตัวอย่าง

3. ส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation)

ส่วนเบี่ยงเบนมาตรฐาน คือ ผลรวมของทุกค่าที่ห่างจากค่ากลางของข้อมูล $\left(X - \bar{X}\right)$ ที่ยกกำลังสองหารด้วยจำนวนข้อมูลลบด้วย 1 แล้วนำค่าที่ได้มาหาค่ารากที่สอง ใช้วิเคราะห์ตัวแปรประเภทมาตรวัดส่วนช่วง (Interval Scale) หรือมาตรวัดส่วนอัตราส่วน (Ratio Scale) ซึ่งมีสูตรคำนวณดังสมการที่ 2.22

$$S.D. = \sqrt{\frac{\left(\sum_{i=1}^{n} X_i - \bar{X}\right)^2}{n-1}} \qquad (2.22)$$

เมื่อ $S.D$ คือ ค่าส่วนเบี่ยงเบนมาตรฐาน
x_i คือ คะแนนของตัวอย่างชุดที่ i
n คือ ขนาดตัวอย่าง

2.5.2. สถิติเชิงอนุมาน (Inferential Statistics)

สถิติเชิงอนุมาน เป็นการนำข้อมูลที่เก็บมาได้จากกลุ่มตัวอย่างไปใช้อ้างอิงหรืออธิบายกลุ่มประชากร ได้แก่ การประเมินค่าพารามิเตอร์ในประชากร (Estimation) และการทดสอบสมมุติฐาน (Hypothesis Testing) แบ่งออกเป็น Parametric และ Nonparametric statistics (วีระศักดิ์, 2557) แม้ว่าสถิติเชิงอนุมานจะมีอยู่หลายตัว แต่ในงานวิจัยจะกล่าวถึงเพียงสถิติที่ใช้ในงานวิจัยเท่านั้น ได้แก่ การทดสอบการแจกแจงปรกติ การวิเคราะห์ความแตกต่างของกลุ่มสองกลุ่มด้วยสถิติทดสอบ Mann-Whitney U ซึ่งมีรายละเอียดดังต่อไปนี้

1. การทดสอบการแจกแจงปรกติ (Normality Test)

1) การทดสอบชาพิโร-วิลก์ (Shapiro-Wilk Test)

การทดสอบชาพิโร-วิลก์เป็นสถิติทดสอบที่ใช้ทดสอบการแจกแจงของตัวแปรเชิงปริมาณ การทดสอบชาพิโร-วิลก์ใช้ในกรณีที่ไม่ทราบค่าเฉลี่ย หรือค่าความแปรปรวนของประชากรก็ได้ และตัวอย่างมีขนาดไม่เกิน 50 หน่วย (สายชล, 2563)

ข้อจำกัดเบื้องต้น (Assumption)

ข้อมูลประกอบด้วยตัวอย่างสุ่ม $X_1, ..., X_n$ ที่มาจากประชากรที่มีฟังก์ชันการแจกแจงเหมาะสม F(x) ที่ไม่ทราบค่า

สมมติฐานของการทดสอบ

H_0 : ตัวอย่างจากประชากรที่มีการแจกแจงปรกติ

H_1 : ตัวอย่างจากประชากรที่ไม่ได้มีการแจกแจงปรกติ

สถิติทดสอบ

$$T = \frac{1}{D}\left[\sum_{i=1}^{k} a_i \left(X^{(n-i+1)} - X^{(i)}\right)\right]^2 \quad \text{โดยที่ } D = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \qquad (2.23)$$

เมื่อ $\bar{X}$ คือ ค่าเฉลี่ยของตัวอย่าง
a_i คือ ค่าสัมประสิทธิ์ที่ได้จากการเปิดตารางของชาพิโร-วิลก์

$X^{(i)}$ คือ ค่าสังเกตตัวอย่างแบบเรียงลำดับ

การสรุปผล

ถ้าค่า T มีค่าเข้าใกล้ 1 แสดงว่าตัวอย่างมีการแจกแจงปรกติ แต่ถ้า T มีค่าน้อย กล่าวคือมีค่าน้อยกว่า 1 มาก แสดงว่าไม่ได้มีการแจกแจงปรกติ

2) การทดสอบลิลลีโฟรส์ (Lilliefors Test)

การทดสอบลิลลีโฟรส์เป็นการทดสอบการแจกแจงของประชากรว่าเป็นการแจกแจงปรกติหรือไม่ โดยเป็นวิธีทดสอบที่มีกำลังการทดสอบสูงกว่าการทดสอบอื่น ๆ ซึ่งจะเหมือนกับการทดสอบของคอลโมโกรอฟ-สมีร์นอฟ (Kolmogorov-Smirnov (K-S) Test) แต่การทดสอบลิลลีโฟรส์จะไม่กำหนดค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐานของประชากร จึงต้องประมาณค่า μ ด้วย $\bar{X}$ และประมาณค่า σ ด้วย S การทดสอบนี้จะใช้เมื่อประชากรมีขนาดมากกว่า 50 หน่วย (สุจิตรา, 2564)

สมมติฐานของการทดสอบ

H_0 : ตัวอย่างถูกสุ่มมาจากประชากรที่มีการแจกแจงปรกติ

H_1 : ตัวอย่างถูกสุ่มมาจากประชากรที่ไม่ได้มีการแจกแจงปรกติ

สถิติทดสอบ

$$D = \max\left|F(X) - S(X)\right| \tag{2.24}$$

โดย $F(X)$ = ความน่าจะเป็นสะสมของตัวอย่าง

$S(X)$ = ความน่าจะเป็นสะสมภายใต้สมมติฐานว่าง

การสรุปผล

ปฏิเสธ H_0 ถ้าค่า p-value ของการทดสอบมีค่าน้อยกว่า α ที่กำหนด

2. การทดสอบของแมนน์-วิตนีย์ (The Mann - Whitney Test)

การทดสอบของแมนน์-วิตนีย์ (The Mann – Whitney test) มีชื่อเรียกอีกอย่างหนึ่งว่าการทดสอบวิลคอกสัน (Wilcoxon test) วิธีของวิลคอกสัน (Wilcoxon test) ใช้ในกรณีที่ตัวอย่างมีขนาดเท่ากัน ในขณะที่วิธีแมนน์-วิตนีย์ (The Mann – Whitney test) ใช้ในกรณีที่ตัวอย่างมีขนาดไม่เท่ากัน ในกรณีตัวอย่างขนาดใหญ่สามารถประมาณด้วยการแจกแจงปรกติมาตรฐาน (Z) (สุจิตรา, 2565)

ข้อกำหนดเบื้องต้น

1. ข้อมูลประกอบด้วยตัวอย่างสุ่ม ด้วยค่าสังเกต $x_1, x_2, ..., x_{n_1}$ จากประชากรที่ 1 และตัวอย่างสุ่มอีก 1 ชุด ด้วยค่าสังเกต $x_1, x_2, ..., x_{n_2}$ จากประชากรที่ 2 ซึ่งเป็นอิสระกัน
2. ตัวอย่าง 2 ชุดนี้เป็นอิสระกัน
3. ค่าตัวแปรสุ่มมีค่าต่อเนื่อง
4. มาตราวัดอย่างน้อยเป็นแบบเรียงลำดับ (Ordinal Scale)
5. ฟังก์ชันการแจกแจงของ 2 ประชากร ต่างกันเฉพาะค่ากลาง (ซึ่งนิยมวัดด้วยมัธยฐาน) นั่นคือ ประชากรทั้ง 2 ประชากร ต้องมีการแจกแจงที่เหมือนกัน ต่างกันเฉพาะค่ากลางเท่านั้น

หมายเหตุ ในทางปฏิบัติไม่จำเป็นต้องทราบว่ามีการแจกแจงแบบใด

สมมติฐาน

ถ้าให้ M_X และ M_Y แทนค่ามัธยฐานของประชากรที่ 1 และ 2 ตามลำดับ อาจทำการทดสอบสองหางหรือหางเดียว ได้ดังนี้

$$H_0 : M_X = M_Y$$
$$H_1 : M_X \neq M_Y \quad \text{หรือ}$$
$$H_0 : M_X \geq M_Y$$
$$H_1 : M_X < M_Y \quad \text{หรือ}$$
$$H_0 : M_X \leq M_Y$$
$$H_1 : M_X > M_Y$$

สถิติที่ใช้ทดสอบ

วิธีของ Wilcoxon และ Mann-Whitney Mann-Whitney ได้แสดงความสัมพันธ์ระหว่างสถิติที่ใช้ทดสอบของเขากับของ Wilcoxon พบว่า ถ้าให้ $T = S - \frac{n_1(n_1+1)}{2}$ แล้วค่า T ที่ได้จะมีค่าเท่ากับ U นั่นเอง เมื่อ U = Mann-Whitney U Statistics หลักในการหาอาณาเขตวิกฤต ค่า T ที่มากเกินไปหรือน้อยเกินไปจะทำให้ปฏิเสธ H_0 เพื่อยอมรับ H_1 ดังนั้น สถิติที่ใช้ทดสอบ คือ

$$T = S - \frac{n_1(n_1+1)}{2}$$

เมื่อ S = ผลรวมลำดับที่ของตัวอย่างขนาด n_1 ในข้อมูลรวมทั้งหมดที่เรียงลำดับแล้ว

การตัดสินใจ

ในกรณีการทดสอบสองหาง ปฏิเสธ H_0 ถ้าพบว่าค่า T น้อยเกินไปหรือใหญ่เกินไป

อาณาเขตวิกฤต คือ $T < W_{\alpha/2}$ หรือ $T > W_{1-\alpha/2}$ เมื่อ $W_{1-\alpha/2} = n_1 n_2 - W_{\alpha/2}$

เมื่อเป็นการทดสอบหางเดียว ด้านน้อยกว่า คือ $H_0 : M_X < M_Y$ ปฏิเสธ H_0

เมื่อพบว่า ค่า T น้อยเกินไป

เมื่อเป็นการทดสอบหางเดียว ด้านมากกว่า คือ $H_0 : M_X > M_Y$ ปฏิเสธ H_0

เมื่อพบว่า ค่า T ใหญ่เกินไป

อาณาเขตวิกฤต คือ $T > W_{1-\alpha}$ เมื่อ $W_{1-\alpha} = n_1 n_2 - W_{\alpha}$

3. การทดสอบวิลคอกสัน-แมน-วิทนีย์ (Wilcoxon–Mann–Whitney Test)

วิลคอก (Wilcoxon, 1990) ได้เสนอสถิติทดสอบวิลคอกสัน-แมน-วิทนีย์ (Wilcoxon–Mann–Whitney Test) หรืออาจเรียกอีกชื่อหนึ่งว่า สถิติทดสอบวิลคอกสันแรงค์ซัม (Wilcoxon Ranked Sum Test) ซึ่งเป็นการทดสอบที่ใช้ผลรวมของลำดับ โดยสถิติทดสอบวิลคอกสัน-แมน-วิทนีย์ เป็นสถิติแบบไม่ใช้พารามิเตอร์ที่มีสมบัติการทดสอบใกล้เคียงกับสถิติแบบใช้พารามิเตอร์ ซึ่งมีประสิทธิภาพในการทดสอบสูง จึงเป็นการทดสอบที่เหมาะสำหรับใช้เปรียบเทียบประชากร 2 กลุ่มที่เป็นอิสระกัน และเมื่อข้อมูลอยู่ในมาตรวัดที่ต่ำกว่ามาตราอันตรภาค โดยจะมีความไวต่อการปฏิเสธสมมติฐานหลัก (H_0) ขึ้นอยู่กับอัตราส่วนของขนาดตัวอย่างและความแปรปรวนของประชากร (Wilcox, 2012)

1) ข้อมูลประกอบด้วยค่าสังเกตของตัวอย่างสุ่ม 2 ตัวอย่าง โดยให้ $X_1, X_2, ..., X_m$ เป็น ตัวอย่างสุ่มขนาด m จากประชากรที่ 1 และให้ $Y_1, Y_2, ..., Y_n$ เป็นตัวอย่างสุ่มขนาด n จากประชากร ที่ 2
2) ข้อตกลงเบื้องต้น (1) ตัวอย่างสุ่มทั้งสองเป็นอิสระกัน
   (2) ข้อมูลจากตัวอย่างสุ่มทั้งสองอย่างน้อยอยู่ในมาตราเรียงอันดับ
   (3) ฟังก์ชันการแจกแจงของประชากรทั้งสองประชากรจะแตกต่างกัน เฉพาะพารามิเตอร์ที่เกี่ยวกับตำแหน่งเท่านั้น
3) สมมติฐานการทดสอบ

$$H_0 : P(X > Y) = P(X < Y)$$

$$H_1 : P(X > Y) \neq P(X < Y)$$

4) สถิติทดสอบ

$$WMW = nm + \frac{m(m+1)}{2} - R_X$$

เมื่อ R_X คือ ผลรวมของลำดับในตัวอย่างที่ 1

ซึ่งจะปฏิเสธสมมติฐานหลัก (H_0) ที่ระดับนัยสำคัญ α เมื่อสถิติทดสอบ WMW มีค่าน้อยกว่า $W_{\frac{\alpha}{2}}$ หรือมีค่ามากกว่า $W_{1-\frac{\alpha}{2}}$ เมื่อ $W_{1-\frac{\alpha}{2}}$ คือ ควอนไทล์ที่ $1-\frac{\alpha}{2}$ ได้จากการเปิดตารางวิลคอกสัน-แมน-วิทนีย์

กรณีที่ตัวอย่างมีขนาดใหญ่ $(m,n \geq 20)$ ภายใต้สมมติฐานหลัก (H_0) เป็นจริง สถิติทดสอบ WMW สามารถประมาณได้ด้วยการแจกแจงปกติมาตรฐานที่มีค่าเฉลี่ย $\frac{mn}{2}$ และความแปรปรวน $\frac{mn(m+n+1)}{12}$ จะได้

สถิติทดสอบ $WMW_1 = \frac{WMW - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}}$

ซึ่งจะปฏิเสธสมมติฐานหลัก (H_0) ที่ระดับนัยสำคัญ α เมื่อสถิติทดสอบ WMW_1 มีค่าน้อยกว่า $z_{\frac{\alpha}{2}}$ หรือมีค่ามากกว่า $z_{1-\frac{\alpha}{2}}$ เมื่อ $z_{1-\frac{\alpha}{2}}$ คือ ควอนไทล์ที่ $1-\frac{\alpha}{2}$ ได้จากการเปิดตารางการแจกแจงแบบปรกติมาตรฐาน

2.6 ระบบพิสูจน์ตัวตน Azure Active Directory

Azure Active Directory (Azure AD) เป็นบริการการตรวจสอบ และการจัดการสำหรับควบคุมการเข้าถึงระบบ Cloud (Cloud-Based Identity And Access Management Service) ซึ่งเป็นส่วนหนึ่งของพื้นที่คลาวด์ของ Microsoft Azure โดยจะช่วยให้พนักงานสามารถเข้าถึง และใช้งานบริการต่าง ๆ ที่พัฒนาโดย Microsoft ได้ง่ายขึ้น Azure AD มีความสามารถในการจัดกับการรับรองตัวตน และการเข้าถึงทรัพยากรขององค์กรอย่างปลอดภัย โดยสามารถใช้งานร่วมกับ Active Directory (AD) ซึ่งเป็นเครื่องมือการจัดการรับรองตัวตนและการเข้าถึงของ Windows Server ได้ ซึ่งจะช่วยให้พนักงานสามารถเข้าถึง และใช้งานทรัพยากรต่าง ๆ ขององค์กรได้โดยไม่ต้องเข้าสู่ระบบโดเมน (On-Premises) ขององค์กร แต่สามารถเข้าถึงผ่านเครือข่ายอินเทอร์เน็ต (Internet) ซึ่งช่วยให้พนักงานภายในองค์กรสามารถทำ Single Sign-on ระหว่างระบบใน Data Center และระบบ Cloud ได้อย่างต่อเนื่องซึ่งช่วยลดการป้อนข้อมูลการรับรองตัวตนซ้ำซ้อน และเพิ่มความสะดวกสบายในการเข้าถึงและใช้งานแอพพลิเคชั่นต่าง ๆ หลังจากที่พนักงานภายในองค์กรทำการพิสูจน์ตัวตนเข้าสู่ระบบผ่าน AD ใน Data Center แล้วจะสามารถเข้าใช้งานแอปพลิเคชันต่าง ๆ บนระบบ Cloud ไม่ว่าจะเป็น Office 365, Dynamic CRM Online, SalesForce.com หรือ Dropbox ได้ทันที โดยไม่ต้องพิสูจน์ตัวตนซ้ำอีกครั้ง (Pattana, 2023)

ข้อดีของ Azure Active Directory (ปาริชาติ, 2564)

1) รองรับการพิสูจน์ตัวตนแบบ Multi-factor Authentication เสริมความแข็งแกร่งในการตรวจสอบผู้ใช้งานก่อนเข้าถึงแอปพลิเคชั่น

2) Self-service Password Management และ Self Service Group Management สำหรับให้ผู้ใช้สามารถรีเซ็ตรหัสผ่านและบริหารจัดการกลุ่มของตนได้ด้วยตัวเอง
3) รองรับการทำงานร่วมกับ Cloud Applications ที่พัฒนาขึ้นมาเอง เพื่อให้จัดการเรื่อง SSO และสิทธิ์ในการใช้งานได้ผ่านทาง SAML 2.0, WS-* Protocol, OpenID และ OAuth
4) ให้บริการภายใต้โครงข่ายมาตรฐานสูงของ Microsoft โดยรองรับ SLA ที่ 99.9%

2.7 ภาษาโปรแกรมที่เกี่ยวข้อง

2.7.1 ภาษาไพธอน (Python Programming Language)

ภาษาไพธอนเป็นภาษาระดับสูงภาษาหนึ่งที่ถูกนำมาใช้งานอย่างกว้างขวาง เนื่องจากเป็นภาษาที่เรียบง่ายสามารถเรียนรู้ และเข้าใจได้อย่างรวดเร็ว นอกจากนี้ไพธอนยังสามารถใช้งานได้ครอบคลุมลักษณะงานที่หลากหลาย รองรับการทำงานบนอุปกรณ์ได้หลายรูปแบบ ทำให้ไพธอนเป็นภาษาโปรแกรมที่ได้รับความนิยมมากที่สุดภาษาหนึ่ง (สุดา, 2563) ทั้งนี้ไพธอนเป็นภาษาที่ถูกพัฒนาขึ้นมาโดยไม่ยึดติดกับแพลตฟอร์ม สามารถรันได้ทั้งระบบ Unix, Linux, Windows 2000, Windows XP หรือแม้แต่ระบบ FreeBSD และไพธอนเป็น Opensource เหมือน PHP ทำให้ทุกคนสามารถนำไพธอนมาพัฒนาโปรแกรมโดยไม่มีค่าใช้จ่ายใด

Data Science กับ Python ซึ่ง Data Science คือ วิทยาศาสตร์ข้อมูล คือศาสตร์ที่รวมเอาความรู้ด้านการเขียนโปรแกรม (Programming) ด้านคณิตศาสตร์ (Mathematics) และด้านสถิติ (Statistics) มาประยุกต์รวมกันเพื่อทำให้ข้อมูลที่มีอยู่เกิดความรู้ใหม่ๆ เกิดเป็นข้อมูลที่มีค่าสามารถนำไปใช้ช่วยสนับสนุนการตัดสินใจวางแผนธุรกิจ และช่วยสร้างประโยชน์ทางธุรกิจได้ซึ่งในการนำข้อมูลผ่านกระบวนการวิเคราะห์ข้อมูลเพื่อนำไปประยุกต์ใช้กับธุรกิจมีความสำคัญอย่างมาก โดยเฉพาะในการตัดสินใจสำหรับการบริหารงานในเรื่องต่าง ๆ ในการวางกลยุทธ์ให้สำหรับองค์กร ดังนั้นในกระบวนการทำ Data Science จึงต้องอาศัยการเขียนโปรแกรมที่นำมาใช้แล้วเกิดความยืดหยุ่น และมีฟังก์ชันการทำงานเพื่อรองรับการประมวลผลทางคณิตศาสตร์ได้ดี ซึ่งภาษาที่นิยมเป็นอย่างตอนนี้ คือ Python (อรพิน, 2564)

ข้อดีของการใช้ Python ทำงาน Data Science สามารถสรุปได้ดังนี้ (อรพิน, 2564)

1) เป็นภาษาที่ง่ายต่อการเรียนรู้สามารถประมวลได้โดยเขียนโปรแกรมเพียงไม่กี่บรรทัดเมื่อเทียบกับภาษาอื่น เช่น ภาษา R
2) ทำงานได้เร็วกว่าภาษา R และ MATLAB
3) มีการจัดการทรัพยากรได้ดีทำให้หน่วยความจำน้อยในการประมวลผลโดยเฉพาะอย่างยิ่งในการทำงานกับข้อมูลขนาดใหญ่
4) มีไลบรารีต่าง ๆ มากมายในด้าน Data Science เช่น NumPy Panda SciPy เป็นต้น

ไลบรารีพื้นฐานสำหรับทำ Data Science

1) Pandas เป็นไลบรารีที่ถูกนำไปใช้ทั้งในเชิงการวิเคราะห์ด้านการเงินเศรษฐกิจ สถิติ รวมถึงการโฆษณา และการประชาสัมพันธ์ ซึ่งมีขั้นตอนในการทำงานที่สำคัญ 5 ขั้นตอน คือ 1) Load 2) Organize 3) Manipulate 4) Model 5) Analyze

2) NumPy ย่อมาจาก Numeric Python ในการใช้งานจะต้องทำการติดตั้งไลบรารีเพิ่มเนื่องจากไม่ใช่ไลบรารีพื้นฐานของไพธอน ซึ่ง NumPy เป็นไลบรารีที่จัดการเกี่ยวกับคณิตศาสตร์ และการคำนวณต่าง ๆ ทั้งข้อมูลที่เกี่ยวกับวิทยาศาสตร์ วิศวกรรมศาสตร์ สถิติ ธุรกิจ เป็นต้น โดยมีความสามารถในการจัดการอาร์เรย์หลายมิติ ซึ่งจะเรียกอาร์เรย์ใน NumPy ว่า ndarray มักใช้คู่กับ SciPy และMatplotlib เพื่อใช้แทน MATLAB

3) Matplotlib ย่อมาจาก Plotting Library เป็นไลบรารีที่ใช้สำหรับการทำ Data Visualization โดยจะมีฟังก์ชั่นเกี่ยวกับการวาดกราฟต่าง ๆ โมดูลที่สำคัญของไลบรารีคือ pyplot ซึ่งเป็นโมดูลหลักของการกำหนดรูปแบบของเส้นกราฟรูปแบบแกนของกราฟ เป็นต้น นอกจากใช้งานร่วมกับ NumPy ยังสามารถใช้งานร่วมกับไลบรารีอื่นที่เกี่ยวกับการวาดกราฟฟิกอย่าง PyQt และ wxPython ได้อีกด้วย

4) Seaborn เป็นไลบรารีสำหรับการพล็อตกราฟในการทำ Data Visualization ซึ่งเรียกใช้งานไลบรารี Matplotlib ในการทำงานอีกทีหนึ่ง มีประโยชน์ในการแสดงกราฟการกระจายของข้อมูลเนื่องจากมีฟังก์ชั่นโดยเฉพาะสำหรับการรับข้อมูลแบบอาร์เรย์ไปแสดงการกระจายชุดข้อมูลในเชิงสถิติ เช่น กราฟแบบระฆังคว่ำ ไคสแควร์ เป็นต้น รวมถึงแสดงกราฟเส้นหรือกราฟแท่ง

2.7.2 ภาษาเอสคิวแอล (Structure Query Language: SQL)

ภาษาที่ใช้ในการจัดการข้อมูลในฐานข้อมูล สามารถแบ่งตามลักษณะการทำงานได้ 4 ส่วนคือ

1) การจัดการเกี่ยวกับการกำหนดโครงสร้างข้อมูล (Data Definition Language: DDL) มีหน้าที่ในการกำหนดโครงสร้างของข้อมูลที่ผู้ใช้แต่ละคนมองเห็น โครงสร้างข้อมูลที่นักออกแบบฐานข้อมูลมองเห็น และโครงสร้างฐานข้อมูลที่จัดเก็บในอุปกรณ์เก็บข้อมูลซึ่งผลของการแปล DDL จะเก็บในไฟล์พิเศษเรียกว่าพจนานุกรมของข้อมูล (Data Dictionary) ตัวอย่างได้แก่ คำสั่ง CREATE, ALTER, DROP, TRUNCATE เป็นต้น

2) การจัดการเกี่ยวกับข้อมูล (Data Manipulation Language: DML) มีหน้าที่ ในด้านจัดการการเข้าถึงข้อมูล ได้แก่ การสอบถามหรือค้นหาข้อมูล (Select) ที่อยู่ในฐานข้อมูล การเพิ่มเติมข้อมูลใหม่ (Insert) เข้าไปในฐานข้อมูล การลบข้อมูล (Delete) ออกจากฐานข้อมูล การเปลี่ยนแปลงแก้ไขข้อมูล (Update) ที่อยู่ในฐานข้อมูล ตัวอย่างคำสั่งได้แก่ SELECT, INSERT, DELETE, UPDATE เป็นต้น

3) การจัดการเกี่ยวกับการประเมินผลกลุ่มงาน (Transaction Processing) เป็นคำสั่งที่เกี่ยวกับการควบคุมประมวลผลกลุ่มงานตาม Business Process ของระบบงานได้แก่ คำสั่ง COMMIT, ROLLBACK, SAVEPOINT

4) การจัดการเกี่ยวกับการใช้สิทธิข้อมูล (Authority) ในฐานข้อมูล เป็นคำสั่งที่เกี่ยวกับการบริหารสิทธิการใช้ข้อมูล (Authority) ของ Database User ในฐานข้อมูล ได้แก่คำสั่ง GRANT, REVOKE

2.8 งานวิจัยที่เกี่ยวข้อง

Zhangyu et al. (2019) ได้ศึกษาการตรวจจับสิ่งผิดปรกติ ด้วยวิธีไอโซเลชั่นฟอเรส และวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ เนื่องจากวิธีไอโซเลชั่นฟอเรสมีความอ่อนแอต่อการจัดการกับค่านอกเกณฑ์ในพื้นที่ (Local Outlier) ในขณะที่วิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ทำงานได้ดีในการตรวจจับค่านอกเกณฑ์ในพื้นที่ แต่ก็มีความซับซ้อนและใช้เวลาในการประมวลผลนาน ผู้วิจัยจึงเสนอวิธีการเรียนรู้แบบรวมกลุ่มผสม (Two-Layer Progressive Ensemble Method) ได้แก่ วิธีไอเอฟ-แอลโอเอฟเพื่อเปรียบเทียบค่าความแม่น (Accuracy) และค่าประสิทธิภาพโดยรวม (F-Measure) กับวิธีไอโซเลชั่นฟอเรส และวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ จากการศึกษาพบว่า วิธีไอเอฟ-แอลโอเอฟมีค่าความแม่น และค่าประสิทธิภาพโดยรวมสูงที่สุด

Goldstein and Uchida (2016) ได้ศึกษาเกี่ยวกับการเปรียบเทียบอัลกอริทึม สำหรับการตรวจจับสิ่งผิดปรกติแบบไม่มีผู้สอนสำหรับข้อมูลหลายตัวแปร (A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data) การตรวจจับสิ่งผิดปรกติเป็นกระบวนการในการระบุรายการ หรือเหตุการณ์ที่เบี่ยงเบนจากชุดข้อมูล โดยการตรวจจับ สิ่งผิดปรกติที่ไม่มีผู้สอนมักคำนึงถึงโครงสร้างภายในของชุดข้อมูลเท่านั้น ซึ่งการตรวจจับสิ่งผิดปรกติแบบไม่มีผู้สอนจะใช้งานจริงในการตรวจจับการบุกรุกเครือข่าย การตรวจจับการทุจริต มีการเสนอ อัลกอริทึมจำนวนมากที่ใช้ในการตรวจจับสิ่งผิดปรกติ แต่ในการวิจัยยังขาดการประเมินที่เป็นสากลในการเปรียบเทียบ ข้อบกพร่องเหล่านี้ได้รับการกล่าวถึงในการศึกษาครั้งนี้ซึ่งใช้อัลกอริทึมการตรวจจับสิ่งผิดปรกติแบบไม่มีผู้สอนที่แตกต่างกัน ได้รับการประเมินจากชุดข้อมูล 10 ชุดที่แตกต่างกัน โดยงานวิจัยมีวัตถุประสงค์เพื่อเป็นพื้นฐานสำหรับการวิจัยตรวจจับสิ่งผิดปรกติที่ไม่มีผู้สอน นอกจากนี้การประเมินผลจะแสดงให้เห็นถึงประสิทธิภาพ จุดอ่อน และจุดแข็งของวิธีการต่าง ๆ รวมถึงพฤติกรรมการตรวจจับสิ่งผิดปรกติ และผลกระทบของการตั้งค่าพารามิเตอร์

Zhang et al. (2007) ได้ศึกษาวิธีการตรวจจับสิ่งผิดปรกติในข้อมูลเครือข่ายการสื่อสาร โดยใช้ วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง (One Class Support Vector Machine for Anomaly Detection in the Communication Network Performance Data) การตรวจจับสิ่งผิดปรกติคือการระบุพฤติกรรมที่ผิดปรกติจากข้อมูลจำนวนมาก โดยการตรวจจับสิ่งผิดปรกติเป็นสิ่งที่จำเป็นมากขึ้น ในเครือข่ายการสื่อสารเนื่องจากจำนวนกิจกรรมที่ไม่ได้รับอนุญาตในเครือข่ายเพิ่มขึ้น บทความนี้ได้นำเสนอ วิธีการที่ใช้คือวิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง เพื่อตรวจจับสิ่งผิดปรกติของเครือข่าย นำข้อมูลเครือข่ายการสื่อสารมาใช้สำหรับการตรวจสอบ และผลลัพธ์ที่ได้ยังถูกนำไปเปรียบเทียบกับ ผลลัพธ์ที่ได้จากวิธีฐานกฎที่ใช้ในปัจจุบันกับเครือข่ายการสื่อสาร พบว่าวิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่งแสดงประสิทธิภาพ และมีแนวโน้มในการตรวจจับสิ่งผิดปรกติของเครือข่ายดีกว่าวิธีฐานกฎ

Niyagas et al. (2003) นำเสนองานวิจัยในเชิงการวิเคราะห์ข้อมูลเชิงธุรกิจ โดยทำการแบ่งกลุ่มข้อมูลลูกค้าที่ใช้งานอินเตอร์เน็ตแบงค์กิ้ง ซึ่งทำการเปรียบเทียบอัลกอริทึมที่ใช้ใน การแบ่งกลุ่มระหว่างอัลกอริทึมเคมีน (K-Means Algorithm) และอัลกอริทึมระบบโครงข่ายประสาทเทียม (Neural Networks Algorithm) โดยข้อมูลที่ใช้ทำงานวิจัยมาจากคลังข้อมูลของลูกค้าที่ใช้อินเตอร์เน็ตแบงค์ เช่น การทำรายการการชำระค่าบริการการโอนเงิน หรือรายการต่าง ๆ ผลที่ได้จากงานวิจัยทำให้เกิดประโยชน์ทางด้านการวางแผนในการรักษาฐานลูกค้าเดิม และการเพิ่มลูกค้าใหม่ซึ่งเป็นการนำข้อมูลดังกล่าวมาใช้ให้เกิดประโยชน์และไม่ทำให้เกิดการสูญเสีย โอกาสในการทำ

ธุรกิจ และพบว่า K-Means Algorithm แบ่งกลุ่มได้ดีกว่าเมื่อมีจำนวนข้อมูล มากในทางกลับกัน Neural Networks Algorithm แบ่งกลุ่มได้ดีกว่าเมื่อข้อมูลมีจำนวนน้อย

Forsmark (2020) ได้ทำการตรวจจับสิ่งผิดปรกติในบันทึกการตรวจสอบสิทธิ์ผู้ใช้โดยใช้ LSTM และ Word Embeddings และได้ทำการคัดเลือกตัวแปรที่มาใช้ในงานวิจัยทั้งหมด 3 ชุด คือ ชุดที่ 1 ผู้ใช้ ช่วงเวลาของวัน ช่วงเวลาเช้า [06.00-12.00] ช่วงเวลากลางวัน [12.00-18.00] ช่วงเวลาเย็น [18.00-00.00] ช่วงเวลาดึก [00.00-06.00] และสถานะการเข้าสู่ระบบสำเร็จ/ล้มเหลว ชุดที่ 2 ผู้ใช้ ประเทศ และสถานะการเข้าสู่ระบบสำเร็จ/ล้มเหลว ชุดที่ 3 ผู้ใช้ ประเทศ และอุปกรณ์ที่ใช้งาน ผลลัพธ์ที่ได้ พบว่าตัวแปรใน ชุดที่ 1 ให้ค่าความแม่นสูงที่สุด

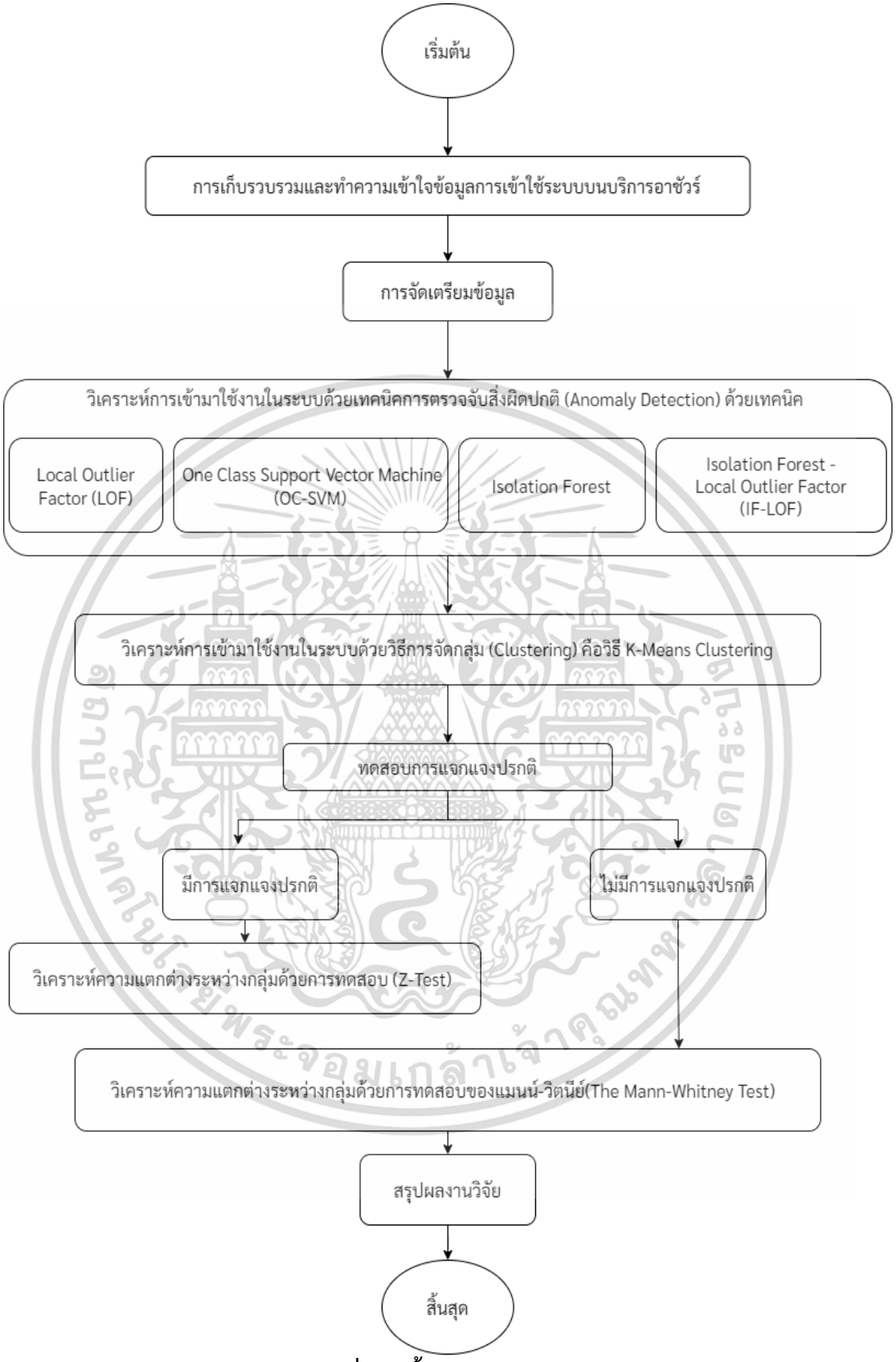
Henriksson (2021) ได้ทำการตรวจจับสิ่งผิดปรกติด้วยเทคนิคการเรียนรู้แบบไม่มีผู้สอนในข้อมูลอนุกรมเวลาโดยใช้ วิธี DBSCAN วิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ วิธีไอโซเลชั่นฟอเรส และวิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง จากทั้ง 4 วิธี พบว่าวิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่งทำงานได้ดีที่สุดในการตรวจจับสิ่งผิดปรกติของจุด ในขณะที่วิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ทำได้ดีที่สุดในการตรวจจับสิ่งผิดปรกติโดยรวม

บทที่ 3
วิธีการดำเนินงานวิจัย

การศึกษาเรื่องการตรวจจับสิ่งผิดปรกติของการเข้าใช้ระบบ Azure Active Directory ด้วยเทคนิคการเรียนรู้ของเครื่อง จากนั้นวิเคราะห์คุณลักษณะของข้อมูลสิ่งผิดปรกติที่ได้อาศัยการจัดกลุ่มด้วยเทคนิคเคมีน และสถิติเชิงอนุมาน เพื่อทำให้เกิดความเข้าใจถึงพฤติกรรมและลักษณะของสิ่งที่ผิดปรกติของการเข้ามาใช้งานในระบบ โดยรายละเอียดการดำเนินงานวิจัยสามารถอธิบายแยกเป็นหัวข้อย่อย ๆ ดังนี้

3.1 ขั้นตอนการดำเนินงานวิจัย

รูปที่ 3.1 แสดงให้เห็นถึงขั้นตอนการดำเนินงานโดยเริ่มจากการเก็บรวบรวม และทำความเข้าใจข้อมูลการเข้าใช้ระบบบน Azure Active Directory โดยมี 2 ขั้นตอน คือ 1) การนำข้อมูลเข้า 2) การทำความเข้าใจข้อมูล จากนั้นทำการจัดเตรียมข้อมูลทั้งหมด 9 ขั้นตอน เพื่อที่จะเก็บจำนวนการเข้ามาใช้งานของระบบในแต่ละช่วงเวลาทั้งหมด 4 ช่วงเวลา ได้แก่ ช่วงเวลาเช้า [06.00-11.59] ช่วงเวลากลางวัน [12.00-17.59] ช่วงเวลาเย็น [18.00-23.59] และช่วงเวลาดึก [00.00-05.59] (Forsmark, 2020) หลังจากการจัดเตรียมข้อมูล ก็ได้นำข้อมูลที่ได้จัดเตรียมไปวิเคราะห์การเข้ามาใช้งานในระบบด้วยเทคนิคการตรวจจับสิ่งผิดปรกติ (Anomaly Detection) ด้วย วิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ (Local Outlier Factor: LOF) วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง (One-Class Support Vector Machine: OC-SVM) วิธีไอโซเลชั่นฟอเรส (Isolation Forest: IF) และวิธีไอเอฟ-แอลโอเอฟ (Isolation Forest- Local Outlier Factor: IF-LOF) จากนั้นวิเคราะห์สิ่งผิดปรกติที่ได้โดยอาศัยวิธีการจัดกลุ่ม (Clustering) ด้วยเทคนิควิธี K-Means Clustering และได้ทำการทดสอบการแจกแจงปรกติพบว่าข้อมูลมีการแจกแจงไม่ปรกติ เมื่อไม่เป็นตามข้อกำหนดเบื้องต้นของสถิติอิงพารามิเตอร์ จึงเลือกใช้การทดสอบของแมนน์-วิตนีย์ (The Mann - Whitney Test) ซึ่งเป็นสถิติไม่อิงพารามิเตอร์ จากนั้นทำการสรุปผลการวิจัย เพื่ออธิบายพฤติกรรมและคุณลักษณะของสิ่งที่ผิดปรกติของการเข้ามาใช้งานระบบในแต่ละช่วงเวลา



รูปที่ 3.1 ขั้นตอนดำเนินงาน

3.2 เครื่องมือที่ใช้ในการวิจัย

3.2.1 ซอฟต์แวร์ที่ใช้ในการวิจัย (Software)

- Microsoft SQL Server Management Studio เป็นระบบการจัดการฐานข้อมูลที่ใช้สำหรับเก็บรวบรวมข้อมูลที่ต้องการศึกษา
- โปรแกรม Visual Studio Code เวอร์ชัน 1.75.1 เป็นโปรแกรมที่ใช้สำหรับสร้างเทคนิคการตรวจจับสิ่งผิดปรกติ
- โปรแกรม IBM SPSS 26 เป็นโปรแกรม สำเร็จรูปที่ใช้สำหรับการวิเคราะห์หาข้อมูลทางสถิติ

3.2.2 ฮาร์ดแวร์ที่ใช้ในการวิจัย (Hardware)

หน่วยประมวลผล AMD Ryzen 7 3750H with Radeon Vega Mobile Gfx .30 GHz หน่วยความจำ 8.0 GB

3.2.3 ชุดคำสั่งที่ใช้ในการวิจัย (Library)

pandas	Pandas คือหนึ่งในชุดคำสั่งสำคัญของภาษา Python มีความสามารถในการจัดการ และวิเคราะห์ข้อมูลได้อย่างมีประสิทธิภาพตั้งแต่ข้อมูลขนาดเล็กไปจนถึงข้อมูลขนาดใหญ่ สามารถใช้การเขียนโค้ดเพื่อปรับแต่ง หรือเชื่อมต่อกับโปรแกรมอื่น ๆ เพื่อดู Data set
pyodbc	เป็นไลบรารี (Library) ในภาษา Python ที่ช่วยในการเชื่อมต่อฐานข้อมูลจาก Python ไปยังฐานข้อมูลต่าง ๆ ได้อย่างสะดวกและง่ายดาย โดย Pyodbc สามารถเชื่อมต่อกับฐานข้อมูลได้หลายชนิด เช่น Microsoft SQL Server, Oracle, MySQL, PostgreSQL, SQLite และอื่น ๆ
os	ในภาษา Python ที่ใช้ในการจัดการและควบคุมระบบไฟล์และโปรแกรมในระบบปฏิบัติการ (Operating System) จัดการ Path: สามารถดึง Path ทั้งหมดของไฟล์ และโฟลเดอร์ในระบบได้ อีกทั้งยังสามารถเพิ่มหรือลบ Path ได้
datetime	ในภาษา Python ที่ช่วยในการจัดการเกี่ยวกับวันที่ และเวลา สามารถแปลงข้อมูลวันที่ และเวลาเป็นรูปแบบที่ต้องการ เช่น แปลงจากวันที่แบบภาษาอังกฤษ (Datetime Object) เป็นวันที่แบบเลข (String) หรือแปลงจากเลขเป็น Datetime object
matplotlib	Matplotlib เป็นชุดคำสั่งของภาษา Python เพื่อใช้ในการสร้างหรือแสดงผล Data visualization ช่วยในการสร้างแผนภูมิ และกราฟต่าง ๆ เพื่อช่วยในการวิเคราะห์ที่ทำให้ดูง่ายขึ้น

sklearn	Scikit-Learn หรือ sklearn เป็นชุดคำสั่งของภาษา Python ใช้สำหรับการเรียนรู้ของเครื่องและสร้างตัวแบบทางสถิติ การจำแนกประเภท เช่น Regression, Classification และ Clustering

3.3 การเก็บรวบรวมและทำความเข้าใจข้อมูลการเข้าใช้ระบบบนบริการอาชัวร์

ผู้วิจัยได้ดำเนินการรวบรวมข้อมูลการเข้ามาใช้งานระบบอาชัวร์ ของบริษัทผลิตเครื่องดื่มแห่งหนึ่งจาก Log File ในฐานข้อมูล ตั้งแต่วันที่ 9 มกราคม พ.ศ. 2566 ถึง วันที่ 16 มีนาคม พ.ศ. 2566

3.3.1 การนำข้อมูลเข้า

ขั้นตอนการนำชุดข้อมูลจากฐานข้อมูลเข้าโปรแกรม Visual Studio Code โดยใช้ชุดคำสั่ง pyodbc เพื่อเชื่อมต่อกับฐานข้อมูล และใช้ชุดคำสั่ง pandas ในการอ่านไฟล์จากฐานข้อมูล

3.3.2 การทำความเข้าใจข้อมูล

หลังจากนำข้อมูลจากฐานข้อมูลของบริษัทมาเก็บไว้ในฐานข้อมูลจำลองเพื่อศึกษารายละเอียดของข้อมูลก่อนนำไปสร้างตัวแบบสำหรับการตรวจจับสิ่งผิดปรกติ โดยผู้วิจัยได้นำข้อมูลการเข้าใช้งานข้อมูลของผู้ใช้งานจากฐานข้อมูลของบริษัทมีรายละเอียด แสดงดังตารางที่ 3.1

ตารางที่ 3.1 รายละเอียดของข้อมูลที่ได้จาก Log Microsoft 365

ข้อมูล	**รายละเอียด**
date_timestamp	วันที่และเวลาที่ทำรายการ
audit_Workload	ชื่อแอปพลิเคชัน
event_provider	ชื่อแอปพลิเคชั่น
audit_UserId	รหัสผู้ใช้งาน
audit_ActorIpAddress	IP address ต้นทาง O365
audit_ClientIP	ClientIP ต้นทาง O365
source_ip	IP Address ต้นทาง
country_iso_code	ชื่อประเทศ
audit_Operation	สถานะการเข้าใช้งาน
event_outcome	สถานะการเข้าใช้งาน
audit_ResultStatus	สถานะการเข้าใช้งาน O365
audit_LogonError	ประเภทการเข้าใช้งานที่เข้าสู่ระบบไม่ถูกต้อง
mfa_status	สถานะปัจจุบัน
audit_Extended_RequestType	ประเภทการขอเข้าใช้งาน
audit_Extended_ResultStatusDetail	สถานะประเภทการขอเข้าใช้งาน
user_agent_original	อุปกรณ์การขอเข้าใช้งาน
event_id	รหัสการเข้าใช้งาน

3.4 การจัดเตรียมข้อมูล

3.4.1 การสกัดข้อมูล (Data Extraction)

ขั้นตอนการสกัดข้อมูลที่จำเป็นจากฐานข้อมูลของบริษัท เพื่อใช้สำหรับการวิเคราะห์และหาวิธีสำหรับการตรวจจับสิ่งผิดปรกติ ซึ่งการสกัดข้อมูลจำเป็นต้องทำความเข้าใจข้อมูลก่อนเพื่อตัดข้อมูลที่ไม่จำเป็น โดยการตรวจจับสิ่งผิดปรกติเน้นการตรวจสอบในสถานะการเข้ามาใช้งานของพนักงานในบริษัท โดยข้อมูลชุดที่ 1 เป็นข้อมูลที่เก็บเฉพาะสถานะการเข้ามาใช้งานในระบบสำเร็จ จำนวน 1,000,000 รายการ ดังตารางที่ 3.2 และข้อมูลชุด ที่ 2 เป็นข้อมูลที่เก็บเฉพาะสถานะการเข้ามาใช้งานในระบบล้มเหลว 253,005 รายการ ดังตารางที่ 3.3

ตารางที่ 3.2 ข้อมูลสถานะการเข้ามาใช้งานในระบบสำเร็จ

ข้อมูล	รายละเอียด
date_timestamp	วันที่และเวลาที่ทำรายการ
audit_UserId	รหัสผู้ใช้งาน
audit_Operation	สถานะการเข้าใช้ระบบสำเร็จ

	date_timestamp	user_email	audit_Operation
0	2023-01-16 08:07:08	AAAAA	UserLoggedIn
1	2023-01-23 07:54:30	BBBBB	UserLoggedIn
2	2023-01-16 07:57:18	CCCCC	UserLoggedIn
3	2023-01-16 07:57:40	AAAAA	UserLoggedIn
4	2023-03-13 08:02:44	CCCCC	UserLoggedIn
...	...	...	...
999995	2023-03-02 04:12:19	DDDDD	UserLoggedIn
999996	2023-02-20 07:51:19	EEEEE	UserLoggedIn
999997	2023-03-03 02:46:34	FFFFF	UserLoggedIn
999998	2023-02-15 03:13:27	FFFFF	UserLoggedIn
999999	2023-02-21 06:48:08	FFFFF	UserLoggedIn

รูปที่ 3.2 ตัวอย่างข้อมูลสถานะการเข้ามาใช้งานในระบบสำเร็จ

จากรูปที่ 3.2 ประกอบด้วย ตัวแปรวันที่และเวลาที่ทำรายการ (date_timestamp) รหัสผู้ใช้งาน (audit_UserId) และสถานะการเข้าใช้งานสำเร็จ (audit_Operation:UserLoggedIn)

ตารางที่ 3.3 ข้อมูลสถานะการเข้ามาใช้งานในระบบล้มเหลว

ข้อมูล	รายละเอียด
date_timestamp	วันที่และเวลาที่ทำรายการ
audit_UserId	รหัสผู้ใช้งาน
audit_Operation	สถานะการเข้าใช้ระบบล้มเหลว

	date_timestamp	user_email	audit_Operation
0	2023-01-09 03:29:11	AAAAA	UserLoginFailed
1	2023-01-09 03:36:11	BBBBB	UserLoginFailed
2	2023-01-09 03:36:49	AAAAA	UserLoginFailed
3	2023-01-09 03:34:48	CCCCC	UserLoginFailed
4	2023-01-09 03:37:29	DDDDD	UserLoginFailed
...	...	...	...
253000	2023-03-15 23:48:04	EEEEE	UserLoginFailed
253001	2023-03-15 23:48:13	EEEEE	UserLoginFailed
253002	2023-03-15 23:49:19	EEEEE	UserLoginFailed
253003	2023-03-15 23:50:27	EEEEE	UserLoginFailed
253004	2023-03-15 23:50:46	EEEEE	UserLoginFailed

รูปที่ 3.3 ตัวอย่างข้อมูลสถานะการเข้ามาใช้งานในระบบล้มเหลว

จากรูปที่ 3.3 ประกอบด้วย ตัวแปรวันที่และเวลาที่ทำรายการ (date_timestamp) รหัสผู้ใช้งาน(audit_UserId) และสถานะการเข้าใช้งานล้มเหลว (audit_Operation: UserLoginFailed)

3.4.2 การแปลงข้อมูล (Data Transformations)

ขั้นตอนการแปลงข้อมูลให้พร้อมสำหรับนำไปใช้ในการวิเคราะห์ ในขั้นตอนนี้จะทำหลังจากการสกัดข้อมูลที่จำเป็น โดยนำข้อมูลไปประมวลผลบน Jupyter notebook เพื่อทำการแปลงข้อมูลให้อยู่ในรูปที่สามารถนำไปวิเคราะห์ต่อได้ โดยมีขั้นตอนดังนี้

ขั้นตอนที่ 1 ทำการแปลงข้อมูลวันที่ และเวลาที่ทำรายการ ให้อยู่ในรูปแบบ date, hour, time โดยมีวัตถุประสงค์เพื่อให้สามารถนับจำนวนการเข้ามาใช้งานในแต่ละวันได้ โดยมีชุดคำสั่งแสดงรายละเอียดดังตารางที่ 3.4 และ 3.5 ดังนี้

ตารางที่ 3.4 ชุดคำสั่งการแปลงข้อมูลวันที่และเวลาที่ทำรายการสถานะสำเร็จ

#Convert timestamp to data hour time	
1	AzureActiveDirectory_Login['date_timestamp'] = pd.to_datetime(AzureActiveDirectory_Login['date_timestamp'])
2	AzureActiveDirectory_Login['date'] = AzureActiveDirectory_Login['date_timestamp'].dt.date
3	AzureActiveDirectory_Login['hour']= AzureActiveDirectory_Login['date_timestamp'].apply(lambda x: x.hour)
4	AzureActiveDirectory_Login['time'] = AzureActiveDirectory_Login['date_timestamp'].dt.strftime('%H:%M:%S')

	date_timestamp	user_email	audit_Operation	date	hour	time
0	2023-01-16 15:07:08	AAAAA	UserLoggedIn	2023-01-16	15	15:07:08
1	2023-01-23 14:54:30	BBBBB	UserLoggedIn	2023-01-23	14	14:54:30
2	2023-01-16 14:57:18	CCCCC	UserLoggedIn	2023-01-16	14	14:57:18
3	2023-01-16 14:57:40	AAAAA	UserLoggedIn	2023-01-16	14	14:57:40
4	2023-03-13 15:02:44	CCCCC	UserLoggedIn	2023-03-13	15	15:02:44
...	...	...	...	...	...	...
999996	2023-02-20 14:51:19	DDDDD	UserLoggedIn	2023-02-20	14	14:51:19
999997	2023-03-03 09:46:34	EEEEE	UserLoggedIn	2023-03-03	9	09:46:34
999998	2023-02-15 10:13:27	EEEEE	UserLoggedIn	2023-02-15	10	10:13:27
999999	2023-02-21 13:48:08	EEEEE	UserLoggedIn	2023-02-21	13	13:48:08

รูปที่ 3.4 ตัวอย่างชุดข้อมูลหลังจากแปลงวันที่และเวลาที่ทำรายการสถานะสำเร็จ

จากรูปที่ 3.4 ข้อมูลที่นำมาจากฐานข้อมูลการเข้ามาใช้งานในระบบสถานะสำเร็จ โดยทำการเลือกมาทั้งหมด 3 ตัวแปร คือ ตัวแปร date_timestamp ตัวแปร audit_UserId และตัวแปร audit_Operation(UserLoggedIn) ซึ่งนำตัวแปร date_timestamp มาเปลี่ยนรูปแบบโดยสร้างคอลัมน์ date เพื่อเก็บข้อมูลวันที่ สร้างคอลัมน์ hour เพื่อเก็บข้อมูลรายชั่วโมง และสร้างคอลัมน์ time เพื่อเก็บข้อมูลเวลา

ตารางที่ 3.5 ชุดคำสั่งการแปลงข้อมูลวันที่และเวลาที่ทำรายการสถานะล้มเหลว

#Convert timestamp to data hour time	
1	AzureActiveDirectory_Logfail ['date_timestamp'] = .pd.to_datetime(AzureActiveDirectory_Logfail ['date_timestamp'])
2	AzureActiveDirectory_Logfail ['date'] = AzureActiveDirectory_Logfail ['date_timestamp'].dt.date
3	AzureActiveDirectory_Logfail['hour']= AzureActiveDirectory_Logfail ['date_timestamp'].apply(lambda x: x.hour)
4	AzureActiveDirectory_ Logfail ['time'] = AzureActiveDirectory_Logfail ['date_timestamp'].dt.strftime('%H:%M:%S')

	date_timestamp	user_email	audit_Operation	date	hour	time
0	2023-01-09 10:29:11	AAAAA	UserLoginFailed	2023-01-09	10	10:29:11
1	2023-01-09 10:36:11	BBBBB	UserLoginFailed	2023-01-09	10	10:36:11
2	2023-01-09 10:36:49	AAAAA	UserLoginFailed	2023-01-09	10	10:36:49
3	2023-01-09 10:34:48	CCCCC	UserLoginFailed	2023-01-09	10	10:34:48
4	2023-01-09 10:37:29	DDDDD	UserLoginFailed	2023-01-09	10	10:37:29
...	...	...	...	...	...	...
253001	2023-03-16 06:48:13	EEEEE	UserLoginFailed	2023-03-16	6	06:48:13
253002	2023-03-16 06:49:19	EEEEE	UserLoginFailed	2023-03-16	6	06:49:19
253003	2023-03-16 06:50:27	EEEEE	UserLoginFailed	2023-03-16	6	06:50:27
253004	2023-03-16 06:50:46	EEEEE	UserLoginFailed	2023-03-16	6	06:50:46

รูปที่ 3.5 ตัวอย่างชุดข้อมูลหลังจากแปลงวันที่และเวลาที่ทำรายการสถานะล้มเหลว

จากรูปที่ 3.5 ข้อมูลที่นำมาจากฐานข้อมูลการเข้ามาใช้งานในระบบสถานะล้มเหลว โดยทำการเลือกมาทั้งหมด 3 ตัวแปร คือ ตัวแปร date_timestamp ตัวแปร audit_UserId และตัวแปร audit_Operation(UserLoggedIn) ซึ่งนำตัวแปร date_timestamp มาเปลี่ยนรูปแบบโดยสร้างคอลัมน์ date เพื่อเก็บข้อมูลวันที่ สร้างคอลัมน์ hour เพื่อเก็บข้อมูลรายชั่วโมง และสร้างคอลัมน์ time เพื่อเก็บข้อมูลเวลา

ขั้นตอนที่ 2 ทำการสร้างคอลัมน์ขึ้นมา 1 คอลัมน์เพื่อที่จะเก็บจำนวนครั้งของการเข้ามาใช้งานในระบบ โดยมีชุดคำสั่งแสดงรายละเอียดดังตารางที่ 3.6 และ 3.7 ดังนี้

ตารางที่ 3.6 ชุดคำสั่งการสร้างคอลัมน์เพื่อเก็บจำนวนครั้งของการเข้ามาใช้งานสถานะสำเร็จ

#Build DataFrame	
1	AzureActiveDirectory_Login['count_login'] = 0

	date_timestamp	user_email	audit_Operation	count_login	date	hour	time
0	2023-01-16 15:07:08	AAAAA	UserLoggedIn	0	2023-01-16	15	15:07:08
1	2023-01-23 14:54:30	BBBBB	UserLoggedIn	0	2023-01-23	14	14:54:30
2	2023-01-16 14:57:18	CCCCC	UserLoggedIn	0	2023-01-16	14	14:57:18
3	2023-01-16 14:57:40	AAAAA	UserLoggedIn	0	2023-01-16	14	14:57:40
4	2023-03-13 15:02:44	CCCCC	UserLoggedIn	0	2023-03-13	15	15:02:44
...	...	...	...	...	...	...	...
999996	2023-02-20 14:51:19	DDDDD	UserLoggedIn	0	2023-02-20	14	14:51:19
999997	2023-03-03 09:46:34	EEEEE	UserLoggedIn	0	2023-03-03	9	09:46:34
999998	2023-02-15 10:13:27	EEEEE	UserLoggedIn	0	2023-02-15	10	10:13:27
999999	2023-02-21 13:48:08	EEEEE	UserLoggedIn	0	2023-02-21	13	13:48:08

รูปที่ 3.6 ตัวอย่างชุดข้อมูลหลังจากสร้างคอลัมน์เพื่อเก็บจำนวนครั้งของการเข้ามาใช้งานสถานะสำเร็จ

จากรูปที่ 3.6 ได้ทำการสร้างคอลัมน์ count login เพื่อทำการเก็บจำนวนครั้งของการเข้ามาใช้งานในระบบสถานะสำเร็จของแต่ละผู้ใช้งาน

ตารางที่ 3.7 ชุดคำสั่งการสร้างคอลัมน์เพื่อเก็บจำนวนครั้งของการเข้ามาใช้งานสถานะล้มเหลว

#Build DataFrame	
1	AzureActiveDirectory_Logfail['count_login'] = 0

	date_timestamp	user_email	audit_Operation	count_login	date	hour	time
0	2023-01-09 10:29:11	AAAAA	UserLoginFailed	0	2023-01-09	10	10:29:11
1	2023-01-09 10:36:11	BBBBB	UserLoginFailed	0	2023-01-09	10	10:36:11
2	2023-01-09 10:36:49	AAAAA	UserLoginFailed	0	2023-01-09	10	10:36:49
3	2023-01-09 10:34:48	CCCCC	UserLoginFailed	0	2023-01-09	10	10:34:48
4	2023-01-09 10:37:29	DDDDD	UserLoginFailed	0	2023-01-09	10	10:37:29
...	...	...	...	...	...	...	...
253001	2023-03-16 06:48:13	EEEEE	UserLoginFailed	0	2023-03-16	6	06:48:13
253002	2023-03-16 06:49:19	EEEEE	UserLoginFailed	0	2023-03-16	6	06:49:19
253003	2023-03-16 06:50:27	EEEEE	UserLoginFailed	0	2023-03-16	6	06:50:27
253004	2023-03-16 06:50:46	EEEEE	UserLoginFailed	0	2023-03-16	6	06:50:46

รูปที่ 3.7 ตัวอย่างชุดข้อมูลหลังจากสร้างคอลัมน์เพื่อเก็บจำนวนครั้งของการเข้ามาใช้งานสถานะล้มเหลว

จากรูปที่ 3.7 ได้ทำการสร้างคอลัมน์ count login เพื่อทำการเก็บจำนวนครั้งของการเข้ามาใช้งานในระบบสถานะล้มเหลวของแต่ละผู้ใช้งาน

ขั้นตอนที่ 3 ทำการรวมกลุ่มคอลัมน์ และนับจำนวนครั้งการเข้ามาใช้งานในแต่ละชั่วโมง โดยมีชุดคำสั่งแสดงรายละเอียดดังตารางที่ 3.8 และ 3.9 ดังนี้

ตารางที่ 3.8 ชุดคำสั่งการรวมกลุ่มคอลัมน์ และนับจำนวนการเข้ามาใช้งานสถานะสำเร็จ

#Groupby column and count value login	
1	AzureActiveDirectory_Login_1H = AzureActiveDirectory_Login.groupby(['date','user_email','audit_Operation','hour']) ['count_login'].count().reset_index()
2	AzureActiveDirectory_Login_1H

	date	user_email	audit_Operation	hour	count_login
0	2023-01-09	AAAAA	UserLoggedIn	21	1
1	2023-01-09	BBBBB	UserLoggedIn	14	3
2	2023-01-09	CCCCC	UserLoggedIn	11	7
3	2023-01-09	CCCCC	UserLoggedIn	13	3
4	2023-01-09	CCCCC	UserLoggedIn	17	4
...	...	...	...	...	...
255461	2023-03-16	DDDDD	UserLoggedIn	5	1
255462	2023-03-16	EEEEE	UserLoggedIn	0	4
255463	2023-03-16	FFFFF	UserLoggedIn	0	6
255464	2023-03-16	GGGGG	UserLoggedIn	0	6
255465	2023-03-16	HHHHH	UserLoggedIn	0	1

รูปที่ 3.8 ตัวอย่างชุดข้อมูลที่ทำการรวมกลุ่ม และนับจำนวนการเข้ามาใช้งานในแต่ละชั่วโมงของข้อมูลสถานะสำเร็จ

จากรูปที่ 3.8 ได้ทำการรวมกลุ่มคอลัมน์ date, user_email, audit_Operation และ hour ของข้อมูลสถานะสำเร็จ หลังจากนั้นนับจำนวนครั้งการเข้ามาใช้งานในแต่ละชั่วโมงเก็บไว้ที่คอลัมน์ count_login จากข้อมูล 1,000,000 รายการ ถูกรวมให้เหลือ 255,466 รายการ

ตารางที่ 3.9 ชุดคำสั่งรวมกลุ่มคอลัมน์และนับจำนวนการเข้ามาใช้งานสถานะล้มเหลว

#Groupby column and count value login	
1	AzureActiveDirectory_Logfail _1H = AzureActiveDirectory_ Logfail.groupby(['date','user_email','audit_Operation','hour']) ['count_login'].count().reset_index()
2	AzureActiveDirectory_Logfail _1H

	date	user_email	audit_Operation	hour	count_login
0	2023-01-09	AAAAA	UserLoginFailed	17	4
1	2023-01-09	BBBBB	UserLoginFailed	14	1
2	2023-01-09	CCCCC	UserLoginFailed	14	4
3	2023-01-09	DDDDD	UserLoginFailed	16	3
4	2023-01-09	EEEEE	UserLoginFailed	13	1
...	...	...	...	...	...
80740	2023-03-16	FFFFF	UserLoginFailed	5	9
80741	2023-03-16	GGGGG	UserLoginFailed	5	6
80742	2023-03-16	HHHHH	UserLoginFailed	6	1
80743	2023-03-16	IIIII	UserLoginFailed	6	4
80744	2023-03-16	JJJJJ	UserLoginFailed	4	101

รูปที่ 3.9 ตัวอย่างชุดข้อมูลที่ทำการจับกลุ่ม และนับจำนวนการเข้ามาใช้งานในแต่ละชั่วโมงของข้อมูลสถานะล้มเหลว

จากรูปที่ 3.9 ได้ทำการรวมกลุ่มคอลัมน์ date ,user_email, audit_Operation และ hour ของข้อมูลสถานะล้มเหลว หลังจากนั้นนับจำนวนครั้งการเข้ามาใช้งานในแต่ละชั่วโมงใส่ไว้ที่คอลัมน์ count_login จากข้อมูล 253,005 รายการ ถูกรวมให้เหลือ 80,745 รายการ

ขั้นตอนที่ 4 ทำ Pivot Table เพื่อที่จะเปลี่ยนรูปแบบคอลัมน์ โดยมีชุดคำสั่งแสดงรายละเอียดดังตารางที่ 3.10 และ 3.11 ดังนี้

ตารางที่ 3.10 ชุดคำสั่งการทำ Pivot Table ของการเข้ามาใช้งานสถานะสำเร็จ

# pivot the table	
1	pivot_AzureActiveDirectory_Login_1H = AzureActiveDirectory_Login_1H.pivot_table(values='count_login', index=['date','user_email'], columns='hour', aggfunc='sum')
2	pivot_AzureActiveDirectory_Login_1H

	hour	0	1	2	3	...	21	22	23
date	user_email								
09-01-2023	AAAAA	NaN	NaN	NaN	NaN	...	1	NaN	NaN
	BBBBB	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
	CCCCC	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
	DDDDD	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
	EEEEE	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...
16-03-2023	FFFFF	NaN	NaN	2	4	...	NaN	NaN	NaN
	GGGGG	4	NaN	NaN	NaN	...	NaN	NaN	NaN
	HHHHH	6	NaN	NaN	NaN	...	NaN	NaN	NaN
	IIIII	6	NaN	NaN	NaN	...	NaN	NaN	NaN
	JJJJJ	1	NaN	NaN	NaN	...	NaN	NaN	NaN

รูปที่ 3.10 ตัวอย่างชุดข้อมูลที่ทำ Pivot Table ของข้อมูลสถานะสำเร็จ

จากรูปที่ 3.10 ได้ทำการ Pivot Table ของข้อมูลสถานะสำเร็จ ซึ่งตารางที่ได้จะมีเค้าโครงแบบใหม่ที่ในแต่ละแถวแสดงถึงชุดค่าผสมของคอลัมน์ date และคอลัมน์ user_email ที่ไม่ซ้ำกัน และคอลัมน์ 0 ถึง คอลัมน์ 23 จะแสดงจำนวนครั้งในการเข้ามาใช้งานระบบในแต่ละรายชั่วโมง

ตารางที่ 3.11 ชุดคำสั่งการทำ Pivot Table ของการเข้ามาใช้งานสถานะล้มเหลว

# pivot the table	
1	pivot_AzureActiveDirectory_Logfail _1H = AzureActiveDirectory_Logfail _1H.pivot_table(values='count_login', index=['date','user_email'], columns='hour', aggfunc='sum')
2	pivot_AzureActiveDirectory_Logfail _1H

	hour	0	1	2	3	...	21	22	23
date	user_email								
09-01-2023	AAAAA	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
	BBBBB	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
	CCCCC	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
	DDDDD	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
	EEEEE	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...
16-03-2023	FFFFF	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
	GGGGG	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
	HHHHH	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
	IIIII	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
	JJJJJ	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN

รูปที่ 3.11 ตัวอย่างชุดข้อมูลที่ทำ Pivot Table ของข้อมูลสถานะล้มเหลว

จากรูปที่ 3.11 ได้ทำการ Pivot Table ของข้อมูลสถานะล้มเหลว ซึ่งตารางที่ได้จะมีเค้าโครงแบบใหม่ที่ในแต่ละแถวแสดงถึงชุดค่าผสมของคอลัมน์ date และคอลัมน์ user_email ที่ไม่ซ้ำกัน และคอลัมน์ 0 ถึง คอลัมน์ 23 จะแสดงจำนวนครั้งในการเข้ามาใช้งานระบบในแต่ละรายชั่วโมง

ขั้นตอนที่ 5 ทำการเปลี่ยนชื่อคอลัมน์ โดยมีชุดคำสั่งแสดงรายละเอียดดังตารางที่ 3.12 และ 3.13 ดังนี้

ตารางที่ 3.12 ชุดคำสั่งการเปลี่ยนชื่อคอลัมน์ของการเข้ามาใช้งานสถานะสำเร็จ

# Rename Column	
1	pivot_AzureActiveDirectory_Login_1H.rename(columns={0: "00.00AM"},inplace=True)
2	pivot_AzureActiveDirectory_Login_1H.rename(columns={1: "01.00AM"},inplace=True)
3	pivot_AzureActiveDirectory_Login_1H.rename(columns={2: "02.00AM"},inplace=True)
4	pivot_AzureActiveDirectory_Login_1H.rename(columns={3: "03.00AM"},inplace=True)
5	pivot_AzureActiveDirectory_Login_1H.rename(columns={4: "04.00AM"},inplace=True)
6	pivot_AzureActiveDirectory_Login_1H.rename(columns={5: "05.00AM"},inplace=True)
7	pivot_AzureActiveDirectory_Login_1H.rename(columns={6: "06.00AM"},inplace=True)
8	pivot_AzureActiveDirectory_Login_1H.rename(columns={7: "07.00AM"},inplace=True)
9	pivot_AzureActiveDirectory_Login_1H.rename(columns={8: "08.00AM"},inplace=True)
10	pivot_AzureActiveDirectory_Login_1H.rename(columns={9: "09.00AM"},inplace=True)
11	pivot_AzureActiveDirectory_Login_1H.rename(columns={10: "10.00AM"},inplace=True)
12	pivot_AzureActiveDirectory_Login_1H.rename(columns={11: "11.00AM"},inplace=True)
13	pivot_AzureActiveDirectory_Login_1H.rename(columns={12: "12.00PM"},inplace=True)
14	pivot_AzureActiveDirectory_Login_1H.rename(columns={13: "13.00PM"},inplace=True)
15	pivot_AzureActiveDirectory_Login_1H.rename(columns={14: "14.00PM"},inplace=True)
16	pivot_AzureActiveDirectory_Login_1H.rename(columns={15: "15.00PM"},inplace=True)
17	pivot_AzureActiveDirectory_Login_1H.rename(columns={16: "16.00PM"},inplace=True)
18	pivot_AzureActiveDirectory_Login_1H.rename(columns={17: "17.00PM"},inplace=True)
19	pivot_AzureActiveDirectory_Login_1H.rename(columns={18: "18.00PM"},inplace=True)
20	pivot_AzureActiveDirectory_Login_1H.rename(columns={19: "19.00PM"},inplace=True)
21	pivot_AzureActiveDirectory_Login_1H.rename(columns={20: "20.00PM"},inplace=True)
22	pivot_AzureActiveDirectory_Login_1H.rename(columns={21: "21.00PM"},inplace=True)
23	pivot_AzureActiveDirectory_Login_1H.rename(columns={22: "22.00PM"},inplace=True)
24	pivot_AzureActiveDirectory_Login_1H.rename(columns={23: "23.00PM"},inplace=True)

	hour	00.00AM	01.00AM	02.00AM	03.00AM	...	21.00PM	22.00PM	23.00PM
date	user_email								
09-01-2023	AAAAA	NaN	NaN	NaN	NaN	...	1	NaN	NaN
	BBBBB	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
	CCCCC	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
	DDDDD	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
	EEEEE	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...
16-03-2023	FFFFF	NaN	NaN	2	4	...	NaN	NaN	NaN
	GGGGG	4	NaN	NaN	NaN	...	NaN	NaN	NaN
	HHHHH	6	NaN	NaN	NaN	...	NaN	NaN	NaN
	IIIII	6	NaN	NaN	NaN	...	NaN	NaN	NaN
	JJJJJ	1	NaN	NaN	NaN	...	NaN	NaN	NaN

รูปที่ 3.12 ตัวอย่างชุดข้อมูลหลังจากที่ทำการเปลี่ยนชื่อคอลัมน์ของข้อมูลสถานะสำเร็จ

จากรูปที่ 3.12 ได้ทำการเปลี่ยนชื่อตัวแปรของข้อมูลสถานะสำเร็จ โดยคอลัมน์ทั้ง 24 คอลัมน์ จากคอลัมน์เลข 0 ถึงเลข 23 เป็นชื่อคอลัมน์ 00.00AM ถึง 23.00PM เพื่อเป็นการศึกษาพฤติกรรมการเข้ามาใช้งานระบบในแต่ละช่วงเวลา (Forsmark, 2020)

ตารางที่ 3.13 ชุดคำสั่งการเปลี่ยนชื่อคอลัมน์ของการเข้ามาใช้งานสถานะล้มเหลว

#	Rename Column
1	pivot_AzureActiveDirectory_Logfail_1H.rename(columns={0: "00.00AM"},inplace=True)
2	pivot_AzureActiveDirectory_Logfail_1H.rename(columns={1: "01.00AM"},inplace=True)
3	pivot_AzureActiveDirectory_Logfail_1H.rename(columns={2: "02.00AM"},inplace=True)
4	pivot_AzureActiveDirectory_Logfail_1H.rename(columns={3: "03.00AM"},inplace=True)
5	pivot_AzureActiveDirectory_Logfail_1H.rename(columns={4: "04.00AM"},inplace=True)
6	pivot_AzureActiveDirectory_Logfail_1H.rename(columns={5: "05.00AM"},inplace=True)
7	pivot_AzureActiveDirectory_Logfail_1H.rename(columns={6: "06.00AM"},inplace=True)
8	pivot_AzureActiveDirectory_Logfail_1H.rename(columns={7: "07.00AM"},inplace=True)
9	pivot_AzureActiveDirectory_Logfail_1H.rename(columns={8: "08.00AM"},inplace=True)
10	pivot_AzureActiveDirectory_Logfail_1H.rename(columns={9: "09.00AM"},inplace=True)
11	pivot_AzureActiveDirectory_Logfail_1H.rename(columns={10: "10.00AM"},inplace=True)
12	pivot_AzureActiveDirectory_Logfail_1H.rename(columns={11: "11.00AM"},inplace=True)
13	pivot_AzureActiveDirectory_Logfail_1H.rename(columns={12: "12.00PM"},inplace=True)
14	pivot_AzureActiveDirectory_Logfail_1H.rename(columns={13: "13.00PM"},inplace=True)
15	pivot_AzureActiveDirectory_Logfail_1H.rename(columns={14: "14.00PM"},inplace=True)
16	pivot_AzureActiveDirectory_Logfail_1H.rename(columns={15: "15.00PM"},inplace=True)
17	pivot_AzureActiveDirectory_Logfail_1H.rename(columns={16: "16.00PM"},inplace=True)
18	pivot_AzureActiveDirectory_Logfail_1H.rename(columns={17: "17.00PM"},inplace=True)
19	pivot_AzureActiveDirectory_Logfail_1H.rename(columns={18: "18.00PM"},inplace=True)
20	pivot_AzureActiveDirectory_Logfail_1H.rename(columns={19: "19.00PM"},inplace=True)
21	pivot_AzureActiveDirectory_Logfail_1H.rename(columns={20: "20.00PM"},inplace=True)

# Rename Column	
22	pivot_AzureActiveDirectory_Logfail _1H.rename(columns={21: "21.00PM"},inplace=True)
23	pivot_AzureActiveDirectory_Logfail _1H.rename(columns={22: "22.00PM"},inplace=True)
24	pivot_AzureActiveDirectory_Logfail _1H.rename(columns={23: "23.00PM"},inplace=True)

	hour	00.00AM	01.00AM	02.00AM	03.00AM	...	21.00PM	22.00PM	23.00PM
date	user_email								
09-01-2023	AAAAA	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
	BBBBB	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
	CCCCC	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
	DDDDD	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
	EEEEE	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...
16-03-2023	FFFFF	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
	GGGGG	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
	HHHHH	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
	IIIII	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
	JJJJJ	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN

รูปที่ 3.13 ตัวอย่างชุดข้อมูลหลังจากที่ทำการเปลี่ยนชื่อคอลัมน์ของข้อมูลสถานะล้มเหลว

จากรูปที่ 3.13 ได้ทำการเปลี่ยนชื่อตัวแปรของข้อมูลสถานะล้มเหลว โดยคอลัมน์ทั้ง 24 คอลัมน์ จากคอลัมน์เลข 0 ถึงเลข 23 เป็นชื่อคอลัมน์ 00.00 AM - 23.59 PM เพื่อเป็นการศึกษาพฤติกรรมการเข้ามาใช้งานระบบในแต่ละช่วงเวลา (Forsmark, 2020)

ขั้นตอนที่ 6 ทำการเติมค่าข้อมูลว่าง หรือ NaN ด้วยตัวเลข 0 โดยมีชุดคำสั่งแสดงรายละเอียดดังตารางที่ 3.14 และ 3.15 ดังนี้

ตารางที่ 3.14 ชุดคำสั่งการเติมค่าข้อมูลว่างด้วยตัวเลข 0 ของการเข้ามาใช้งานสถานะสำเร็จ

#fill NaN	
1	pivot_AzureActiveDirectory_Login_1H = pivot_AzureActiveDirectory_Login_1H.fillna(value=0)
2	pivot_AzureActiveDirectory_Login_1H

	hour	00.00AM	01.00AM	02.00AM	03.00AM	...	21.00PM	22.00PM	23.00PM
date	user_email								
09-01-2023	AAAAA	0	0	0	0	...	1	0	0
	BBBBB	0	0	0	0	...	0	0	0
	CCCCC	0	0	0	0	...	0	0	0
	DDDDD	0	0	0	0	...	0	0	0
	EEEEE	0	0	0	0	...	0	0	0
...	...	...	...	...	...	...	...	...	...
16-03-2023	FFFFF	0	0	2	4	...	0	0	0
	GGGGG	4	0	0	0	...	0	0	0
	HHHHH	6	0	0	0	...	0	0	0
	IIIII	6	0	0	0	...	0	0	0
	JJJJJ	1	0	0	0	...	0	0	0

รูปที่ 3.14 ตัวอย่างชุดข้อมูลหลังจากการเติมค่าข้อมูลว่างด้วยตัวเลข 0 ของการเข้ามาใช้งานสถานะสำเร็จ

จากรูปที่ 3.14 ได้ทำการเปลี่ยนแปลงข้อมูลของการเข้ามาใช้งานสถานะสำเร็จจากที่ในแต่ละเซลล์เป็น NaN ให้เป็นตัวเลข 0 เพื่อที่จะแทนค่าว่างให้อยู่ในรูปตัวเลข เพื่อที่จะสามารถนำข้อมูลไปทำการวิเคราะห์ต่อได้

ตารางที่ 3.15 ชุดคำสั่งการเติมค่าข้อมูลว่างด้วยตัวเลข 0 ของการเข้ามาใช้งานสถานะล้มเหลว

#fill NaN	
1	pivot_AzureActiveDirectory_Logfail_1H = pivot_AzureActiveDirectory_Logfail_1H.fillna(value=0)
2	pivot_AzureActiveDirectory_Logfail_1H

	hour	00.00AM	01.00AM	02.00AM	03.00AM	...	21.00PM	22.00PM	23.00PM
date	user_email								
09-01-2023	AAAAA	0	0	0	0	...	0	0	0
	BBBBB	0	0	0	0	...	0	0	0
	CCCCC	0	0	0	0	...	0	0	0
	DDDDD	0	0	0	0	...	0	0	0
	EEEEE	0	0	0	0	...	0	0	0
...	...	...	...	...	...	...	...	...	...
16-03-2023	FFFFF	0	0	0	0	...	0	0	0
	GGGGG	0	0	0	0	...	0	0	0
	HHHHH	0	0	0	0	...	0	0	0
	IIIII	0	0	0	0	...	0	0	0
	JJJJJ	0	0	0	0	...	0	0	0

รูปที่ 3.15 ตัวอย่างชุดข้อมูลหลังจากการเติมค่าข้อมูลว่างด้วยตัวเลข 0 ของการเข้ามาใช้งานสถานะล้มเหลว

จากรูปที่ 3.15 ได้ทำการเปลี่ยนแปลงข้อมูลของการเข้ามาใช้งานสถานะล้มเหลวจากที่ในแต่ละเซลล์เป็น NaN ให้เป็นตัวเลข 0 เพื่อที่จะแทนค่าว่างให้อยู่ในรูปตัวเลข เพื่อที่จะสามารถนำข้อมูลไปทำการวิเคราะห์ต่อได้

ขั้นตอนที่ 7 สร้างคอลัมน์เพิ่มขึ้นมา 4 คอลัมน์เพื่อที่ทำการรวมจำนวนครั้งที่เข้ามาใช้งานระบบเวลารายชั่วโมงให้เป็นช่วงเวลา โดยมีชุดคำสั่งแสดงรายละเอียดดังตารางที่ 3.16 และ 3.17 ดังนี้

ตารางที่ 3.16 ชุดคำสั่งการสร้างคอลัมน์ช่วงเวลาของการเข้ามาใช้งานสถานะสำเร็จ

# Build Column To Defind time period	
1	pivot_AzureActiveDirectory_Login_1H['Morning'] = 0 #06.00-11.59
2	pivot_AzureActiveDirectory_Login_1H['Afternoon'] = 0 #12.00-17.59
3	pivot_AzureActiveDirectory_Login_1H['Evening'] = 0 #18.00-23.59
4	pivot_AzureActiveDirectory_Login_1H['Night'] = 0 #00.00-05.59

	hour	00.00AM	01.00AM	...	23.00PM	Morning	Afternoon	Evening	Night
date	user_email								
09-01-2023	AAAAA	0	0	...	0	0	0	0	0
	BBBBB	0	0	...	0	0	0	0	0
	CCCCC	0	0	...	0	0	0	0	0
	DDDDD	0	0	...	0	0	0	0	0
	EEEEE	0	0	...	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...
16-03-2023	FFFFF	0	0	...	0	0	0	0	0
	GGGGG	4	0	...	0	0	0	0	0
	HHHHH	6	0	...	0	0	0	0	0
	IIIII	6	0	...	0	0	0	0	0
	JJJJJ	1	0	...	0	0	0	0	0

รูปที่ 3.16 ตัวอย่างชุดข้อมูลหลังจากการสร้างคอลัมน์ช่วงเวลาของการเข้ามาใช้งานสถานะสำเร็จ

จากรูปที่ 3.16 ได้ทำการสร้างคอลัมน์ Morning [06.00-11.59], Afternoon [12.00-17.59], Evening [18.00-23.59] และ Night [00.00-05.59] ของการเข้ามาใช้งานสถานะสำเร็จเพื่อที่ทำการรวมจำนวนครั้งที่เข้ามาใช้งานระบบเวลารายชั่วโมงให้แบ่งออกเป็น 4 ช่วงเวลา เพื่อนำช่วงเวลาไปวิเคราะห์ต่อไป

ตารางที่ 3.17 ชุดคำสั่งการสร้างคอลัมน์ช่วงเวลาของการเข้ามาใช้งานสถานะล้มเหลว

# Build Column To Defind time period	
1	pivot_AzureActiveDirectory_ Logfail _1H['Morning'] = 0 #06.00-11.59
2	pivot_AzureActiveDirectory_ Logfail _1H['Afternoon'] = 0 #12.00-17.59
3	pivot_AzureActiveDirectory_ Logfail _1H['Evening'] = 0 #18.00-23.59

4	pivot_AzureActiveDirectory_ Logfail _1H['Night'] = 0 #00.00-05.59

	hour	00.00AM	01.00AM	...	23.00PM	Morning	Afternoon	Evening	Night
date	user_email								
09-01-2023	AAAAA	0	0	...	0	0	0	0	0
	BBBBB	0	0	...	0	0	0	0	0
	CCCCC	0	0	...	0	0	0	0	0
	DDDDD	0	0	...	0	0	0	0	0
	EEEEE	0	0	...	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...
16-03-2023	FFFFF	0	0	...	0	0	0	0	0
	GGGGG	0	0	...	0	0	0	0	0
	HHHHH	0	0	...	0	0	0	0	0
	IIIII	0	0	...	0	0	0	0	0
	JJJJJ	0	0	...	0	0	0	0	0

รูปที่ 3.17 ตัวอย่างชุดข้อมูลหลังจากการสร้างคอลัมน์ช่วงเวลาของการเข้ามาใช้งานสถานะล้มเหลว

จากรูปที่ 3.17 ได้ทำการสร้างคอลัมน์ Morning [06.00-11.59], Afternoon [12.00-17.59], Evening [18.00-23.59] และ Night [00.00-05.59] ของการเข้ามาใช้งานสถานะล้มเหลว เพื่อที่ทำการรวมจำนวนครั้งที่เข้ามาใช้งานระบบเวลารายชั่วโมงให้แบ่งออกเป็น 4 ช่วงเวลา เพื่อนำช่วงเวลาไปวิเคราะห์ต่อไป

ขั้นตอนที่ 8 ทำการรวมจำนวนครั้งที่เข้ามาใช้งานระบบในรายชั่วโมงให้เป็นช่วงเวลาเช้า ช่วงเวลากลางวัน ช่วงเวลาเย็น และช่วงเวลาดึก โดยจะเลือกแสดงเฉพาะคอลัมน์ช่วงเวลาซึ่งมีชุดคำสั่งแสดงรายละเอียดดังตารางที่ 3.18 และ 3.19 ดังนี้

ตารางที่ 3.18 ชุดคำสั่งที่ทำการรวมเวลารายชั่วโมงให้เป็นช่วงเวลาสถานะการเข้ามาใช้งานระบบสำเร็จ

#Sum value	
1	pivot_AzureActiveDirectory_Login_1H['Morning'] = pivot_AzureActiveDirectory_Login_1H['06.00AM']+pivot_AzureActiveDirectory_Login_1H['07.00AM']+pivot_AzureActiveDirectory_Login_1H['08.00AM']+pivot_AzureActiveDirectory_Login_1H['09.00AM']+pivot_AzureActiveDirectory_Login_1H['10.00AM']+pivot_AzureActiveDirectory_Login_1H['11.00AM']
2	pivot_AzureActiveDirectory_Login_1H['Afternoon'] = pivot_AzureActiveDirectory_Login_1H['12.00PM']+pivot_AzureActiveDirectory_Login_1H['13.00PM']+pivot_AzureActiveDirectory_Login_1H['14.00PM']+pivot_AzureActiveDirectory_Login_1H['15.00PM']+pivot_AzureActiveDirectory_Login_1H['16.00PM']+pivot_AzureActiveDirectory_Login_1H['17.00PM']
3	pivot_AzureActiveDirectory_Login_1H['Evening'] = pivot_AzureActiveDirectory_Login_1H['18.00PM']+pivot_AzureActiveDirectory_Login_1H['19.00PM']+pivot_AzureActiveDirectory_Login_1H['20.00PM']+pivot_AzureActiveDirectory_Login_1H['

	21.00PM']+pivot_AzureActiveDirectory_Login_1H['22.00PM']+pivot_AzureActiveDirectory_Login_1H['23.00PM']
4	pivot_AzureActiveDirectory_Login_1H['Night'] = pivot_AzureActiveDirectory_Login_1H['00.00AM']+pivot_AzureActiveDirectory_Login_1H['01.00AM']+pivot_AzureActiveDirectory_Login_1H['02.00AM']+pivot_AzureActiveDirectory_Login_1H['03.00AM']+pivot_AzureActiveDirectory_Login_1H['04.00AM']+pivot_AzureActiveDirectory_Login_1H['05.00AM']

	hour	Morning	Afternoon	Evening	Night
date	user_email				
09-01-2023	AAAAA	0	0	1	0
	BBBBB	0	3	0	0
	CCCCC	7	7	1	0
	DDDDD	0	8	0	0
	EEEEE	2	1	0	0
...	...	...	...	...	...
16-03-2023	FFFFF	0	0	0	10
	GGGGG	0	0	0	4
	HHHHH	0	0	0	6
	IIIII	0	0	0	6
	JJJJJ	0	0	0	1

รูปที่ 3.18 ตัวอย่างชุดข้อมูลหลังจากที่ทำการรวมเวลารายชั่วโมงให้เป็นช่วงเวลาของข้อมูลสถานะการเข้ามาใช้งานระบบสำเร็จ

จากรูปที่ 3.18 ได้ทำการทำการรวมเวลารายชั่วโมงให้เป็นช่วงเวลาโดยที่คอลัมน์

- Morning จะเป็นการรวมคอลัมน์ 06.00AM - 11.00AM
- Afternoon จะเป็นการรวมคอลัมน์ 12.00PM - 17.00PM
- Evening จะเป็นการรวมคอลัมน์ 18.00PM - 23.00PM
- Night จะเป็นการรวมคอลัมน์ 00.00AM - 05.00AM

เพื่อที่จะนำข้อมูลไปวิเคราะห์เพื่อดูพฤติกรรมการเข้ามาใช้งานในแต่ละช่วงเวลา จากข้อมูล 255,466 รายการ ถูกรวมให้เหลือ 115,748 รายการ

ตารางที่ 3.19 ชุดคำสั่งทำการรวมเวลารายชั่วโมงให้เป็นช่วงเวลาสถานะการเข้ามาใช้งานระบบล้มเหลว

#Sum value	
1	pivot_AzureActiveDirectory_Logfail_1H['Morning'] = pivot_AzureActiveDirectory_Logfail_1H['06.00AM']+pivot_AzureActiveDirectory_Logfail_1H['07.00AM']+pivot_AzureActiveDirectory_Logfail_1H['08.00AM']+pivot_AzureActiveDirectory_Logfail_1H['09.00AM']+pivot_AzureActiveDirectory_Logfail_1H['10.00AM']+pivot_AzureActiveDirectory_Logfail_1H['11.00AM']

#Sum value	
2	pivot_AzureActiveDirectory_Logfail_1H['Afternoon'] = pivot_AzureActiveDirectory_Logfail_1H['12.00PM']+pivot_AzureActiveDirectory_Logfail_1H['13.00PM']+pivot_AzureActiveDirectory_Logfail_1H['14.00PM']+pivot_AzureActiveDirectory_Logfail_1H['15.00PM']+pivot_AzureActiveDirectory_Logfail_1H['16.00PM']+pivot_AzureActiveDirectory_Logfail_1H['17.00PM']
3	pivot_AzureActiveDirectory_Logfail_1H['Evening'] = pivot_AzureActiveDirectory_Logfail_1H['18.00PM']+pivot_AzureActiveDirectory_Logfail_1H['19.00PM']+pivot_AzureActiveDirectory_Logfail_1H['20.00PM']+pivot_AzureActiveDirectory_Logfail_1H['21.00PM']+pivot_AzureActiveDirectory_Logfail_1H['22.00PM']+pivot_AzureActiveDirectory_Logfail_1H['23.00PM']
4	pivot_AzureActiveDirectory_Logfail_1H['Night'] = pivot_AzureActiveDirectory_Logfail_1H['00.00AM']+pivot_AzureActiveDirectory_Logfail_1H['01.00AM']+pivot_AzureActiveDirectory_Logfail_1H['02.00AM']+pivot_AzureActiveDirectory_Logfail_1H['03.00AM']+pivot_AzureActiveDirectory_Logfail_1H['04.00AM']+pivot_AzureActiveDirectory_Logfail_1H['05.00AM']

	hour	Morning	Afternoon	Evening	Night
date	user_email				
09-01-2023	AAAAA	0	4	0	0
	BBBBB	0	1	0	0
	CCCCC	0	4	0	0
	DDDDD	0	3	0	0
	EEEEE	0	6	0	0
...	...	...	...	...	...
16-03-2023	FFFFF	0	0	0	9
	GGGGG	0	0	0	6
	HHHHH	1	0	0	0
	IIIII	4	0	0	0
	JJJJJ	0	0	0	101

รูปที่ 3.19 ตัวอย่างชุดข้อมูลหลังจากที่ทำการรวมเวลารายชั่วโมงให้เป็นช่วงเวลาของข้อมูลสถานะการเข้ามาใช้งานระบบล้มเหลว

จากรูปที่ 3.19 ได้ทำการทำการรวมเวลารายชั่วโมงให้เป็นช่วงเวลาโดยที่คอลัมน์

- Morning จะเป็นการรวมคอลัมน์ 06.00AM - 11.00AM
- Afternoon จะเป็นการรวมคอลัมน์ 12.00PM - 17.00PM
- Evening จะเป็นการรวมคอลัมน์ 18.00PM - 23.00PM
- Night จะเป็นการรวมคอลัมน์ 00.00AM - 05.00AM

เพื่อที่จะนำข้อมูลไปวิเคราะห์เพื่อดูพฤติกรรมการเข้ามาใช้งานในแต่ละช่วงเวลา จากข้อมูล 80,745 รายการ ถูกรวมให้เหลือ 54,263 รายการ

ขั้นตอนที่ 9 ทำการจัดรูปแบบข้อมูลใหม่เพื่อที่จะไปนำไปตรวจจับสิ่งผิดปรกติในแต่ละช่วงเวลาโดยมีชุดคำสั่งแสดงรายละเอียดดังตารางที่ 3.20 และ 3.21 ดังนี้

ตารางที่ 3.20 ชุดคำสั่งทำการจัดรูปแบบข้อมูลใหม่ของสถานะการเข้ามาใช้งานระบบสำเร็จ

# reset index	
1	rein_pivot_AzureActiveDirectory_Login_1H = pivot_AzureActiveDirectory_Login_1H.reset_index()
2	rein_pivot_AzureActiveDirectory_Login_1H

	date	user_email	Morning	Afternoon	Evening	Night
0	2023-01-09	AAAAA	0.0	0.0	1.0	0.0
1	2023-01-09	BBBBB	0.0	3.0	0.0	0.0
2	2023-01-09	CCCCC	7.0	7.0	1.0	0.0
3	2023-01-09	DDDDD	0.0	8.0	0.0	0.0
4	2023-01-09	EEEEE	2.0	1.0	0.0	0.0
...	...	...	...	...	...	...
115744	2023-03-16	FFFFF	0.0	0.0	0.0	4.0
115745	2023-03-16	GGGGG	0.0	0.0	0.0	6.0
115746	2023-03-16	HHHHH	0.0	0.0	0.0	6.0
115747	2023-03-16	IIIII	0.0	0.0	0.0	1.0

รูปที่ 3.20 ตัวอย่างชุดข้อมูลหลังจากที่ทำการเรียงค่าข้อมูลใหม่ของข้อมูลสถานะการเข้ามาใช้งานระบบสำเร็จ

จากรูปที่ 3.20 ได้ทำจัดรูปแบบข้อมูลใหม่ของข้อมูลสถานะการเข้ามาใช้งานระบบสำเร็จโดยข้อมูลที่ได้จะเป็นจำนวนครั้งของการเข้ามาใช้งานของผู้ใช้ในแต่ละวันโดยที่จะสามารถนำข้อมูลที่ได้ไปตรวจจับสิ่งผิดปรกติ และวิเคราะห์พฤติกรรมการเข้ามาใช้งานระบบในแต่ละช่วงเวลา

ตารางที่ 3.21 ชุดคำสั่งทำการจัดรูปแบบข้อมูลใหม่ของสถานะการเข้ามาใช้งานระบบล้มเหลว

# reset index	
1	rein_pivot_AzureActiveDirectory_Logfail_1H = pivot_AzureActiveDirectory_ Logfail _1H.reset_index()
2	rein_pivot_AzureActiveDirectory_ Logfail _1H

	date	user_email	Morning	Afternoon	Evening	Night
0	2023-01-09	AAAAA	0.0	4.0	0.0	0.0
1	2023-01-09	BBBBB	0.0	1.0	0.0	0.0
2	2023-01-09	CCCCC	0.0	4.0	0.0	0.0
3	2023-01-09	DDDDD	0.0	3.0	0.0	0.0
4	2023-01-09	EEEEE	0.0	6.0	0.0	0.0
...	...	...	...	...	...	...
54259	2023-03-16	FFFFF	0.0	0.0	0.0	6.0
54260	2023-03-16	GGGGG	1.0	0.0	0.0	0.0
54261	2023-03-16	HHHHH	4.0	0.0	0.0	0.0
54262	2023-03-16	IIIII	0.0	0.0	0.0	101.0

รูปที่ 3.21 ตัวอย่างชุดข้อมูลหลังจากที่ทำการเรียงค่าข้อมูลใหม่ของข้อมูลสถานะการเข้ามาใช้งานระบบล้มเหลว

จากรูปที่ 3.21 ได้ทำการจัดรูปแบบข้อมูลใหม่ของข้อมูลสถานะการเข้ามาใช้งานระบบล้มเหลวโดยข้อมูลที่ได้จะเป็นจำนวนครั้งของการเข้ามาใช้งานของผู้ใช้ในแต่ละวันโดยที่จะสามารถนำข้อมูลที่ได้ไปตรวจจับสิ่งผิดปกติ และวิเคราะห์พฤติกรรมการเข้ามาใช้งานระบบในแต่ละช่วงเวลา

3.5 การตรวจจับสิ่งผิดปกติ (Anomaly Detection)

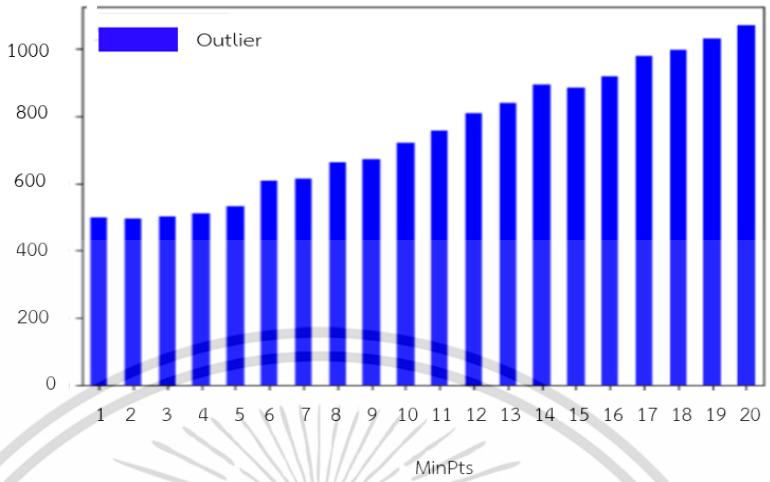
โดยทางผู้วิจัยได้ศึกษาการเรียนรู้แบบไม่มีผู้สอน ซึ่งงานวิจัยของ Henriksson (2021) ได้ใช้การตรวจจับสิ่งผิดปกติด้วยเทคนิคการเรียนรู้แบบไม่มีผู้สอน คือวิธี DBSCAN วิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ (Local Outlier Factor: LOF) วิธีไอโซเลชั่นฟอเรส (Isolation Forest: IF) และวิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง (One-Class Support Vector Machine: OC-SVM) เนื่องจาก วิธี DBSCAN และวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ คือวิธีการหาความหนาแน่นของเพื่อนบ้านเช่นเดียวกัน แต่ผลของวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ให้ประสิทธิภาพในการตรวจจับสิ่งผิดปกติโดยรวมที่ดีกว่า และวิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่งให้ประสิทธิภาพในการตรวจจับสิ่งผิดปกติของจุดได้ดีกว่า และทางผู้วิจัยได้มีศึกษาเพิ่มเติมจากงานวิจัยของ Zhangyu et al. (2019) ที่ใช้วิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ วิธีไอโซเลชั่นฟอเรส และคิดค้นผสมผสานทั้งสองวิธีเป็นวิธีไอเอฟ-แอลโอเอฟ (Isolation Forest- Local Outlier Factor: IF-LOF) และได้ผลลัพธ์คือวิธีไอเอฟ-แอลโอเอฟ มีค่าความแม่นสูงที่สุด ฉะนั้นทางผู้วิจัยเลือกศึกษาวิธีทั้ง 4 วิธี ได้แก่ วิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง วิธีไอโซเลชั่นฟอเรส และวิธีไอเอฟ-แอลโอเอฟ

3.5.1 วิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ (Local Outlier Factor: LOF)

เป็นการตรวจจับสิ่งผิดปกติที่อาศัยความหนาแน่นของเพื่อนบ้านใกล้เคียง โดยผู้วิจัยได้กำหนดพารามิเตอร์ และกระบวนการในการสร้างวิธีการตรวจจับความผิดปกติดังนี้

1. การกำหนดจำนวนเพื่อนบ้านใกล้เคียง (MinPts) โดยเพื่อนบ้านใกล้เคียงอาจมีจำนวนน้อยกว่า หรือเท่ากับจำนวนของข้อมูลในที่นี้ผู้วิจัยได้ทำการรันค่า MinPts ตั้งแต่ 1 ถึง 20

เพื่อเปรียบเทียบดูขั้นตอนวิธีให้จำนวนค่าที่ผิดปรกติในแต่ละ MinPts เท่าใด และได้ทำการ เลือกค่า MinPts ที่ให้ค่าที่ผิดปรกติมากที่สุด



รูปที่ 3.22 จำนวนค่าที่ผิดปรกติของแต่ละ MinPts

2. การเลือกค่าที่ผิดปรกติ จากการทบทวนวรรณกรรมของ (Breunig et al.,2000) ได้เลือกค่าที่ผิดปรกติโดยพิจารณาจากค่าโลคอลเอาท์ไลเออร์แฟคเตอร์ โดยค่าโลคอลเอาท์ไลเออร์แฟคเตอร์ ที่น้อยกว่า -1.5 ถือว่าเป็นค่าที่ผิดปรกติ

จากชุดคำสั่ง sklearn โดยกำหนดจำนวนเพื่อนบ้านใกล้เคียงเป็น 20 แสดงรายละเอียดดังตารางที่ 3.22

ตารางที่ 3.22 ชุดคำสั่งวิธี Local Outlier Factor (LOF)

	# ชุดคำสั่งวิธี Local Outlier Factor
1	from sklearn.neighbors import LocalOutlierFactor
2	ModelLOF = LocalOutlierFactor(n_neighbors=20)
3	y_pred1 = ModelLOF.fit_predict(X1_Std) #predict
4	LOF1 = ModelLOF.negative_outlier_factor_
5	LOF1 = pd.DataFrame(LOF1)
6	LOF1.columns = ["LOF"]
7	y_pred1 = pd.DataFrame(y_pred1)
8	y_pred1["Predict"] = pd.DataFrame(y_pred1)
9	y_pred1 = y_pred1.drop(y_pred1.columns[[0]], axis=1)
10	df_Predictuser = pd.concat([df_user,LOF1,y_pred1], axis=1)
11	df_Predictuser

3.5.2 วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง (One-Class Support Vector Machine: OC-SVM)

เป็นการตรวจจับสิ่งผิดปรกติที่พยายามสร้างขอบเขตการตัดสินใจที่ทำให้เกิดการแยกจากกันของข้อมูล โดยผู้วิจัยได้กำหนดพารามิเตอร์ และกระบวนการในการสร้างวิธีการตรวจจับสิ่งผิดปรกติดังนี้

1. การกำหนดค่า nu (υ) เป็นสัดส่วนของขอบเขตบน และขอบเขตของค่าที่ผิดปรกติ โดยที่ υ มีค่าอยู่ระหว่าง 0 ถึง 1 ซึ่งเป็นพารามิเตอร์ที่กำหนดโดยผู้ใช้ในที่นี้กำหนดเป็น 0.02

2. การกำหนดเคอร์เนล (kernel) เนื่องจากข้อมูลที่ใช้มีรูปแบบที่ไม่สามารถแบ่งได้ด้วยเส้นตรง จึงต้องมีเครื่องมือมาช่วยให้ข้อมูลเรียงตัวใหม่ในพื้นที่ โดยผู้วิจัยเลือกใช้ Radial Basis Function Kernel (RBF) ในการสร้างวิธีการตรวจจับสิ่งผิดปรกติ

3. การกำหนดค่าแกมมา (γ) เป็นค่าสัมประสิทธิ์สำหรับเคอร์เนลโดยกำหนดเป็น “auto” คำนวณได้จาก 1/n_features

4. ค่าคะแนนที่ได้จะบ่งบอกถึงสิ่งผิดปรกติของข้อมูลโดยคำนวณได้จาก Score_Samples(X) จะคำนวณระยะทางจากตัวอย่างไปยังขอบเขตการตัดสินใจของแบบจำลอง แสดงรายละเอียดดังตารางที่ 3.23

ตารางที่ 3.23 ชุดคำสั่งวิธี One-Class Support Vector Machine (OCSVM)

	# ชุดคำสั่งวิธี One-Class Support Vector Machine
1	from sklearn.svm import OneClassSVM # import model OCSVM
2	Modelocsvm= OneClassSVM(nu=0.02,kernel="rbf",gamma=”auto”)
3	Modelocsvm.fit(X1_Std)
4	y_pred1 = Modelocsvm.predict(X1_Std)
5	SVM1 = Modelocsvm.score_samples(X1_Std)
6	SVM1 = pd.DataFrame(SVM1)
7	SVM1.columns = ["SVM"]
8	y_pred1 = pd.DataFrame(y_pred1)
9	y_pred1["Predict"] = pd.DataFrame(y_pred1))
10	y_pred1 = y_pred1.drop(y_pred1.columns[[0]], axis=1)
11	df_Predictuser = pd.concat([df_user,SVM1,y_pred1], axis=1
12	df_Predictuser

3.5.3 วิธีไอโซเลชั่นฟอเรส (Isolation Forest: IF)

วิธีไอโซเลชั่นฟอเรสเป็นเทคนิคการเรียนรู้แบบไม่มีผู้ดูแล ที่ใช้ในการตรวจจับปัญหาที่เกี่ยวข้องกับสิ่งผิดปรกติ (Anomaly Detection) โดยวิธีนี้จะใช้ต้นไม้ (Tree) ในการแบ่งข้อมูล

ออกเป็นกลุ่มย่อย ๆ โดยผู้วิจัยได้กำหนดพารามิเตอร์ และกระบวนการในการสร้างวิธีการตรวจจับสิ่งผิดปรกติดังนี้

1. การกำหนด n_estimators คือการกำหนดจำนวนต้นไม้ตัดสินใจ (Decision Trees) ในกรณีนี้จะตั้งค่าต้นไม้ตัดสินใจเป็น 1,000 หมายความว่าจะมีการสร้างแผนผังการตัดสินใจ 1,000 ต้น

2. การกำหนด max_samples='auto' คือการกำหนดจำนวนตัวอย่างที่จะดึงจากข้อมูลเพื่อฝึกต้นไม้แต่ละต้นค่า 'auto' หมายถึงจำนวนตัวอย่างที่ใช้สำหรับการฝึกอบรมต้นไม้แต่ละต้นจะเท่ากับจำนวนตัวอย่างทั้งหมดในชุดข้อมูล

3. การกำหนด max_features= 4 คือการกำหนดจำนวนคุณลักษณะหรือตัวแปรสูงสุดที่ต้องการพิจารณาเมื่อแยกโหนดในแผนผังการตัดสินใจแต่ละรายการ ในกรณีนี้จะตั้งค่าเป็น 4 เนื่องจากต้องการพิจารณาทุกตัวแปร

4. เมื่อทราบความลึกของต้นไม้แล้วทำให้สามารถคำนวณเป็นคะแนนความผิดปรกติเพื่อใช้ในการแยกประเภทของข้อมูลได้ ซึ่งคะแนนความความผิดปรกตินั้นจะมีค่าน้อยกว่า -0.5 จะหมายถึงข้อมูลที่มีแนวโน้มผิดปรกติ ส่วนข้อมูลที่มีค่ามากกว่า -0.5 จะถือว่าเป็นข้อมูลทั่วไปที่ไม่มีความผิดปรกติ (Hui, 2021)

ตารางที่ 3.24 ชุดคำสั่งวิธี Isolation Forest

# ชุดคำสั่งวิธี Isolation Forest	
1	from sklearn.ensemble import IsolationForest
2	IF_UserLog= IsolationForest (n_estimators=1000, max_samples='auto',max_features=4)
3	IF_UserLog.fit(X_scaled_df_log)
4	predictions_UserLog= IF_UserLog.fit_predict(X_scaled_df_log)
5	predictions_UserLog
# Calculate anomaly scores for each observation	
6	scores = IF_UserLog.score_samples(X_scaled_df_log)
7	scores=pd.DataFrame(scores)
8	concat_UseLog = pd.concat([df_userlog,scores], axis=1)
9	concat_UseLog.rename(columns={0: "scores"},inplace=True)
10	y_pred_UseLog= pd.DataFrame(predictions_UserLog)
11	concat_UseLog = pd.concat([concat_UseLog,y_pred_UseLog], axis=1)

3.5.4 วิธีไอเอฟ-แอลโอเอฟ (Isolation Forest- Local Outlier Factor: IF-LOF)

วิธีไอเอฟ-แอลโอเอฟ (Zhangyu et al, 2019) โดยขั้นตอนการทำงานของวิธีไอเอฟ-แอลโอเอฟ จะทำการนำสิ่งผิดปรกติที่ได้จากการสร้างแบบจำลองวิธีไอโซเลชั่นฟอเรสมาใช้ในการสร้างแบบจำลองวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ โดยผู้วิจัยได้กำหนดพารามิเตอร์ และกระบวนการ

ในการสร้างวิธีการตรวจจับสิ่งผิดปรกติเหมือนกับวิธีไอโซเลชั่นฟอเรส และวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ที่กล่าวไปข้างต้น แสดงรายละเอียดดังตารางที่ 3.25

ตารางที่ 3.25 ชุดคำสั่งวิธี Isolation Forest- Local Outlier Factor (IF-LOF)

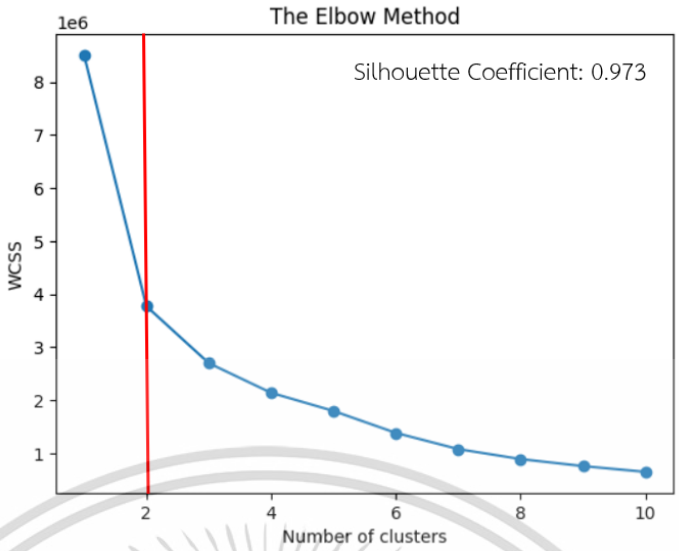
# ชุดคำสั่งวิธี Isolation Forest-Local Outlier Factor	
1	from sklearn.ensemble import IsolationForest
2	from sklearn.neighbors import LocalOutlierFactor
3	IF_UserLog= IsolationForest (n_estimators=1000,max_samples='auto' ,max_features=4)
4	IF_UserLog.fit(X_scaled_df_log)
5	predictions_UserLog= IF_UserLog.fit_predict(X_scaled_df_log)
6	predictions_UserLog
# Calculate anomaly scores for each observation	
7	scores = IF_UserLog.score_samples(X_scaled_df_log)
8	scores=pd.DataFrame(scores)
9	concat_UseLog = pd.concat([df_userlog,scores], axis=1)
10	concat_UseLog.rename(columns={0: "scores"},inplace=True)
11	y_pred_UseLog= pd.DataFrame(predictions_UserLog)
12	concat_UseLog = pd.concat([concat_UseLog,y_pred_UseLog], axis=1)
13	concat_UseLog.rename(columns={0: "Predict"},inplace=True)
14	ModelLOF = LocalOutlierFactor(n_neighbors=20)
15	y_pred1 = ModelLOF.fit_predict(concat_UseLog) #predict
16	y_pred1 = pd.DataFrame(y_pred1)

3.6 การจัดกลุ่มด้วยเทคนิคเคมีน

หลังจากที่ได้สิ่งผิดปรกติจากวิธีการตรวจจับสิ่งผิดปรกติทั้ง 4 วิธี แล้วนั้นนำไปวิเคราะห์ผลเพื่อดูคุณลักษณะของสิ่งผิดปรกติสร้างวิธีการจัดกลุ่มโดยใช้วิธีการจัดกลุ่มแบบไม่เป็นขั้นตอน (K-Means Cluster Analysis) ซึ่งในขั้นตอนการจัดกลุ่ม ผู้วิจัยจะมีการกำหนดค่า K หรือจำนวนกลุ่มด้วยวิธี Elbow วัดจากค่าความคลาดเคลื่อนของผลรวมระยะห่างระหว่างวัตถุ (Object) กับจุดศูนย์กลางเริ่มต้น (Centroid) หรือ ค่าผลบวกกำลังสองภายในกลุ่ม และวิธี Silhouette โดยใช้ค่าเฉลี่ยของระยะห่างระหว่างจุดกับจุดต่าง ๆ ภายในกลุ่มเดียวกัน ส่วนด้วยระยะห่างน้อยที่สุดของจุดกับจุดต่าง ๆ ในแต่ละกลุ่ม เพื่อที่จะหาจำนวนกลุ่ม (K) ที่เหมาะสมที่สุดชุดคำสั่ง sklearn แสดงรายละเอียดดังตารางที่ 3.26 ซึ่งได้ผลลัพธ์ดังรูปที่ดังต่อไปนี้

ตารางที่ 3.26 ชุดคำสั่ง K-Means Clustering

# ชุดคำสั่งวิธี K-Means Clustering **# Using the elbow method to find the optimal number of clusters**	
1	from sklearn.cluster import KMeans
2	from sklearn.metrics import silhouette_score
3	wcss =[]
4	for i in range (1,11):
5	kmeans = KMeans(n_clusters = i, init = 'k-means++', max_iter =300, n_init =10, random_state = 0)
6	kmeans.fit(X_in)
7	wcss.append(kmeans.inertia_)
# Plot the graph to visualize the Elbow Method to find the optimal number of cluster	
8	plt.plot(range(1,11),wcss,marker='o')
9	plt.title('The Elbow Method')
10	plt.xlabel('Number of clusters')
11	plt.ylabel('WCSS')
12	plt.show()
# Applying KMeans to the dataset with the optimal number of cluster	
13	kmeans=KMeans(n_clusters= 3, init = 'k-means++', max_iter = 300, n_init = 10, random_state = 0)
14	y_kmeans = kmeans.fit_predict(X_in)
# Calculate the Silhouette Coefficient	
15	score = silhouette_score(X, kmeans.labels_)



รูปที่ 3.23 แผนภาพตัวอย่าง Elbow หาค่า K ที่เหมาะสม และค่า Silhouette

จากรูปที่ 3.23 จะพบว่าจำนวนกลุ่มที่เหมาะสมในการจัดกลุ่มคือ 2 กลุ่ม (K = 2) และค่า Silhouette Coefficient มีค่าเท่ากับ 0.973 ซึ่งมีค่าใกล้ 1 แสดงว่าข้อมูลถูกจัดให้อยู่ในกลุ่มมีความเหมาะสมแล้ว ซึ่งถ้ามองจากภาพแผนภาพ Elbow ที่มีจุดหักศอกที่ 2 ก็มีความเป็นไปได้ที่ข้อมูลจะแบ่งเป็น 2 กลุ่มเช่นเดียวกัน แสดงว่าการจัดกลุ่ม 2 กลุ่มเหมาะสมกับข้อมูลชุดนี้ที่สุด จึงทำการจัดกลุ่มข้อมูล 2 กลุ่ม ได้ผลลัพธ์เป็นคอลัมน์จัดกลุ่มดังรูปที่ 3.24

	Cluster
0	0
1	0
2	0
3	0
4	0
...	...
1167	1
1168	1
1169	1
1170	1

รูปที่ 3.24 ตัวอย่างคอลัมน์การจัดกลุ่มข้อมูล

3.7 การวิเคราะห์ข้อมูล

3.7.1 สถิติเชิงพรรณนา ใช้สำหรับวิเคราะห์ข้อมูลการเข้ามาใช้งานในแต่ละช่วงเวลาได้แก่ ค่าเฉลี่ย ค่าร้อยละ ส่วนเบี่ยงเบนมาตรฐาน

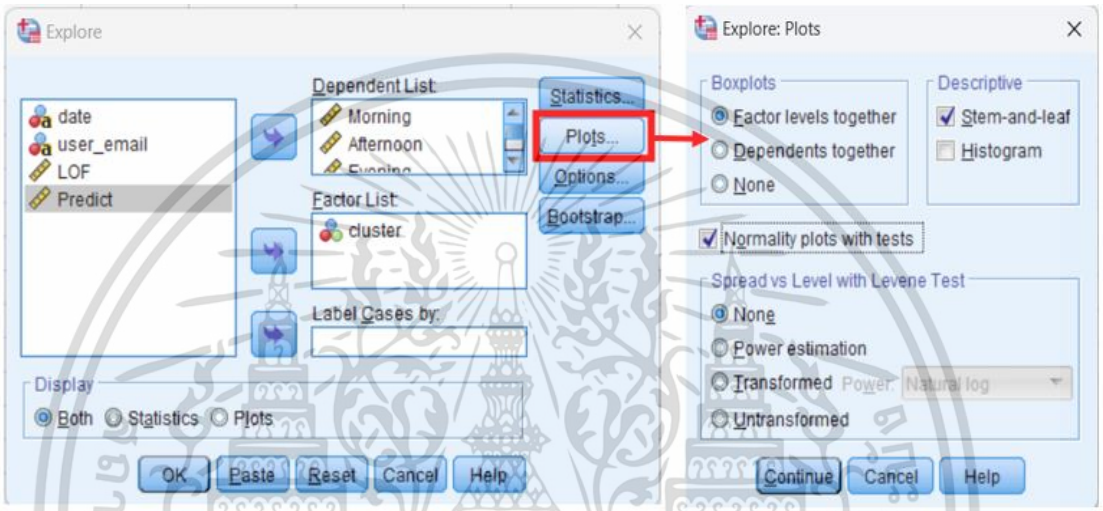
3.7.2 สถิติเชิงอนุมาน การวิเคราะห์ความสัมพันธ์ระหว่างกลุ่มที่ได้จากการจัดกลุ่ม (Clustering) โดยสถิติอนุมาน (inferential Statistics) ด้วยโปรแกรมสำเร็จรูป IBM SPSS มีขั้นตอนดังต่อไปนี้

3.7.2.1 การทดสอบข้อจำกัดเบื้องต้น

เพื่อตรวจสอบข้อมูลว่าข้อมูลมีการแจกแจงปรกติหรือไม่ ถ้าข้อมูลไม่มีการแจกแจงปรกติ ควรจะใช้การวิเคราะห์สถิติแบบไม่ใช้พารามิเตอร์

ขั้นตอนการวิเคราะห์ด้วย SPSS

1) Analyze
2) Descriptive
3) Explore...



รูปที่ 3.25 วิเคราะห์การแจกแจงข้อมูลด้วยโปรแกรมสำเร็จรูป IBM SPSS

Tests of Normality

	cluster	Kolmogorov-Smirnov[a] Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
Morning	0	.401	1765	.000	.193	1765	.000
	1	.241	6	.200*	.903	6	.393
Afternoon	0	.369	1765	.000	.245	1765	.000
	1	.265	6	.200*	.937	6	.639
Evening	0	.441	1765	.000	.089	1765	.000
	1	.243	6	.200*	.892	6	.328
Night	0	.461	1765	.000	.063	1765	.000
	1	.366	6	.012	.636	6	.001

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

รูปที่ 3.26 ผลลัพธ์การวิเคราะห์การแจกแจงข้อมูล

จากรูปที่ 3.26 พบว่าข้อมูลส่วนใหญ่ไม่มีการแจกแจงปรกติ จึงเลือกใช้การใช้สถิติแบบไม่ใช้พารามิเตอร์ (Nonparametric Statistics)

3.7.2.2 การทดสอบสมมติฐาน

เนื่องจากต้องการทดสอบค่ามัธยฐานของจำนวนข้อมูลสิ่งผิดปรกติของช่วงเวลาในแต่ละกลุ่ม โดยใช้การทดสอบของแมนน์-วิตนีย์ (The Mann-Whitney Test) เพื่อตรวจ

ว่าคุณลักษณะ หรือตัวแปรสามารถแบ่งค่าความแตกต่างระหว่างกลุ่มได้อย่างมีนัยสำคัญทางสถิติ กรณีข้อมูลไม่มีการแจกแจงปรกติ

สมมติฐานของการทดสอบ

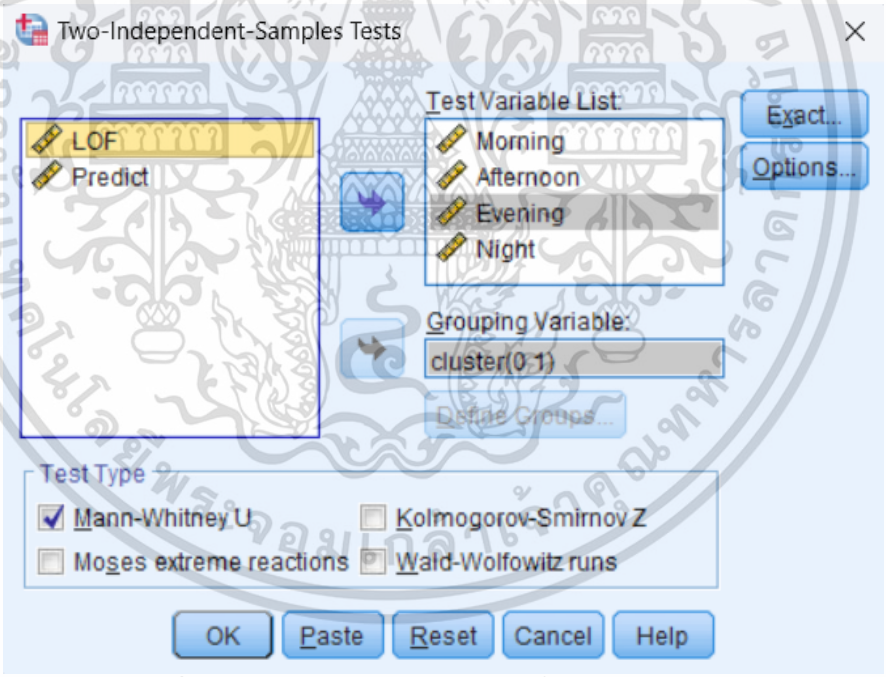
$H_0 : M_X = M_Y$

$H_1 : M_X \neq M_Y$ หรือ

เนื่องจากค่า p-value อยู่ในอาณาเขตวิกฤต จึงปฏิเสธ H_0 จึงสรุปได้ค่ามัธยฐานของจำนวนข้อมูลสิ่งผิดปรกติในกลุ่มที่ 1 และกลุ่มที่ 2 แตกต่างกันที่ระดับนัยสำคัญ 0.05

ขั้นตอนการวิเคราะห์ด้วย SPSS

1) Analyze
2) Nonparametic Test
3) Legacy Dialogs
4) Two Independent Samples...



รูปที่ 3.27 หน้าต่างการเลือกข้อมูลเพื่อใช้ทดสอบความแตกต่าง

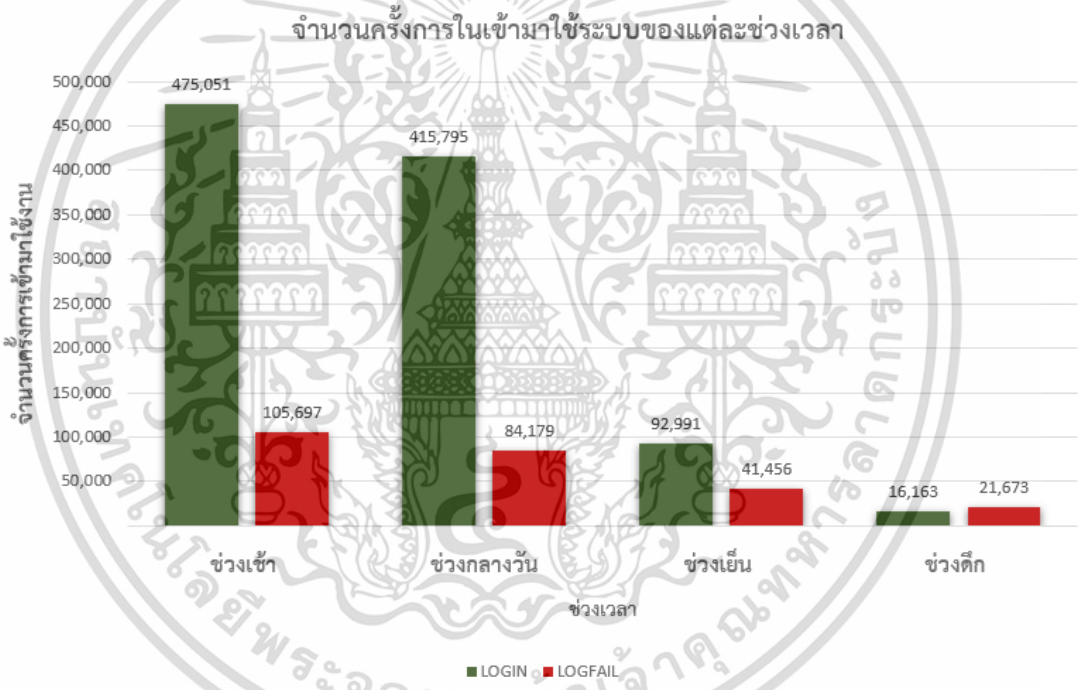
จากรูปที่ 3.27 การทดสอบพบว่าค่ามัธยฐานของจำนวนข้อมูลสิ่งผิดปรกติในกลุ่มที่ 1 และกลุ่มที่ 2 มีความแตกต่างระหว่าง โดยใช้สถิติทดสอบ Mann-Whitney U

บทที่ 4
ผลการวิจัยและอภิปรายผล

ในบทนี้จะกล่าวถึงผลของการตรวจจับสิ่งผิดปรกติที่เกิดขึ้นจากจำนวนรายการที่เข้ามาใช้บริการในระบบ Azure Active Directory ในช่วงเวลาต่าง ๆ ซึ่งแบ่งออกเป็น 4 ช่วงเวลา ได้แก่ ช่วงเวลาเช้า [06.00-11.59] ช่วงเวลากลางวัน [12.00-17.59] ช่วงเวลาเย็น [18.00-23.59] และช่วงเวลาดึก [00.00-05.59] พร้อมทั้งผลการวิเคราะห์ เพื่อทำให้เกิดความเข้าใจถึงพฤติกรรมและลักษณะของสิ่งผิดปรกติ โดยเนื้อหาของบทนี้ประกอบด้วยหัวข้อย่อย ๆ ดังนี้

4.1. ผลการวิเคราะห์ข้อมูลด้วยสถิติพรรณนา

4.1.1. ข้อมูลรายการเข้ามาใช้งานในระบบ



รูปที่ 4.1 จำนวนครั้งในการเข้ามาใช้งานในระบบของแต่ละช่วงเวลา

จากรูปที่ 4.1 แสดงจำนวนครั้งรายการเข้ามาใช้งานในระบบของแต่ละช่วงเวลาทั้งข้อมูลในระบบที่มีสถานะสำเร็จและล้มเหลวจะพบว่า ในช่วงเวลาเช้า [06.00-11.59] จะมีจำนวนรายการค่อนข้างสูงเมื่อเปรียบเทียบกับช่วงเวลาอื่น ๆ โดยคิดเป็นร้อยละ 47.5 สำหรับสถานะสำเร็จ และ 41.7 สำหรับสถานะล้มเหลว

ตารางที่ 4.1 สถิติพรรณนาของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จ

สถิติพรรณนาของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จ							
ตัวแปร	Min	Max	ค่าเฉลี่ย	SD	Median	IQR	Skewness
ช่วงเวลาเช้า	0	1466	4.10	12.83	2	5	49.09
ช่วงเวลากลางวัน	0	2124	3.59	13.97	1	4	67.72
ช่วงเวลาเย็น	0	1449	0.80	9.17	0	0	109.94
ช่วงเวลาดึก	0	1445	0.13	4.98	0	0	222.78

ตารางที่ 4.2 สถิติพรรณนาของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลว

สถิติพรรณนาของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลว							
ตัวแปร	Min	Max	ค่าเฉลี่ย	SD	Median	IQR	Skewness
ช่วงเวลาเช้า	0	669	1.94	9.97	1	2	29.85
ช่วงเวลากลางวัน	0	540	1.55	9.37	0	1	32.76
ช่วงเวลาเย็น	0	598	0.76	8.28	0	0	39.47
ช่วงเวลาดึก	0	611	0.39	7.88	0	0	53.40

จากตารางที่ 4.1 และ 4.2 จะพบว่าค่าเฉลี่ยจำนวนครั้งรายการเข้ามาใช้งานในระบบของแต่ละช่วงเวลาทั้งข้อมูลในระบบที่มีสถานะสำเร็จและล้มเหลวจะพบว่า โดยเฉลี่ยในช่วงเวลาเช้าจะมีจำนวนรายการเข้ามาใช้งานในระบบมากที่สุด ตามด้วยช่วงเวลากลางวัน เย็น และดึกตามลำดับ ซึ่งค่าเฉลี่ยการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จจะสูงกว่าสถานะล้มเหลว เนื่องจากการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลวมีการตั้งค่าว่าถ้ามีการเข้ามาใช้งานในระบบล้มเหลวติดต่อกัน 5 ครั้ง ระบบจะล็อคการเข้ามาใช้งานในระบบ จึงทำให้การเข้ามาใช้งานในระบบที่มีสถานะล้มเหลวมีค่าเฉลี่ยที่ต่ำกว่า นอกจากนี้ยังพบว่าส่วนเบี่ยงเบนมาตรฐานในแต่ละช่วงเวลามีค่าสูง แสดงให้เห็นว่าการกระจายตัวของรายการเข้ามาใช้งานในระบบมากสำหรับสถานะสำเร็จ ในส่วนของสถานะล้มเหลวจะพบว่าในแต่ละช่วงเวลามีการกระจายตัวที่สูงแต่ใกล้เคียงกัน และข้อมูลในระบบที่มีสถานะสำเร็จและล้มเหลวมีลักษณะเบ้ขวาสามารถดูได้จากค่า Skewness ที่คำนวณได้มีค่าเป็นบวก นั่นคือ รายการเข้ามาใช้งานในระบบที่มีจำนวนครั้งน้อยจะมีความถี่ มากกว่ารายการเข้ามาใช้งานในระบบที่มีจำนวนครั้งมาก

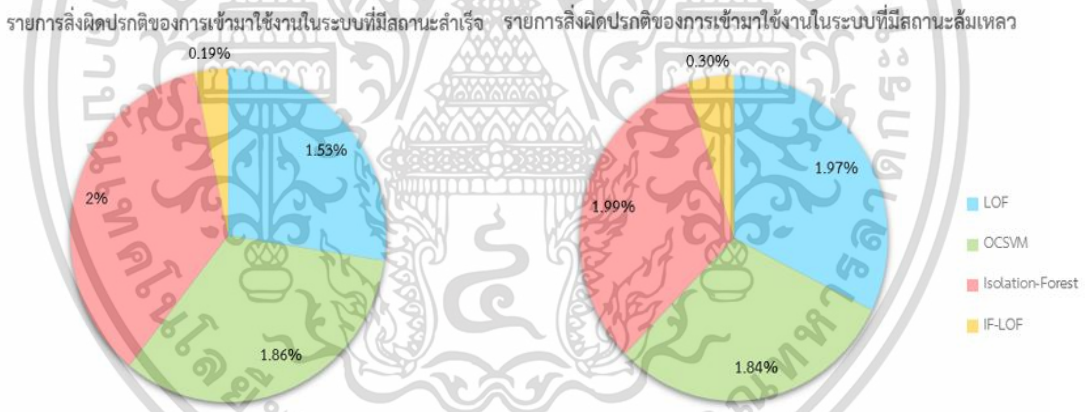
4.2. ผลการตรวจจับสิ่งผิดปรกติทั้ง 4 วิธี

ผลการตรวจจับสิ่งผิดปรกติโดยอาศัยวิธีการตรวจจับสิ่งผิดปรกติ 4 วิธี ได้แก่ วิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ (Local Outlier Factor: LOF) วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง (One-Class Support Vector Machine: OC-SVM) วิธีไอโซเลชั่นฟอเรส (Isolation Forest: IF) และวิธีไอเอฟ-แอลโอเอฟ (Isolation Forest- Local Outlier Factor: IF-LOF) ซึ่งทางผู้วิจัยได้แสดงผลลัพธ์ในการตรวจจับของแต่ละแบบจำลองในข้อมูล 2 ชุดดังนี้

ตารางที่ 4.3 จำนวนและร้อยละของสิ่งผิดปรกติ

ชุดข้อมูล	LOF		OCSVM		Isolation-Forest		IF-LOF	
	จำนวน	%	จำนวน	%	จำนวน	%	จำนวน	%
รายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จ	1,771	1.53	2,160	1.86	2,315	2.00	228	0.19
รายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลว	1,070	1.97	1,003	1.84	1,081	1.99	167	0.30

จากตารางที่ 4.3 แสดงจำนวนและร้อยละสิ่งผิดปรกติ ของรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จและล้มเหลว โดยพบว่าวิธีไอโซเลชั่นฟอเรสสามารถตรวจจับสิ่งผิดปรกติได้ 2,315 รายการคิดเป็นร้อยละ 2 สำหรับสถานะสำเร็จ และ 1,081 รายการคิดเป็นร้อยละ 1.99 สำหรับสถานะล้มเหลว ซึ่งสามารถตรวจจับได้มากที่สุดในขณะที่วิธีไอเอฟ-แอลโอเอฟสามารถตรวจจับสิ่งผิดปรกติได้เพียง 228 รายการคิดเป็นร้อยละ 0.19 สำหรับสถานะสำเร็จ และ 167 รายการคิดเป็นร้อยละ 0.30 สำหรับสถานะล้มเหลว ซึ่งสามารถตรวจจับได้น้อยที่สุด โดยเมื่อนำไปสร้างแผนภูมิรูปวงกลมจะมีลักษณะดังรูปที่ 4.2



รูปที่ 4.2 สัดส่วนรายการสิ่งผิดปรกติที่ตรวจจับได้ของข้อมูลในระบบที่มีสถานะสำเร็จและล้มเหลว

จากรูปที่ 4.2 แสดงสัดส่วนรายการสิ่งผิดปรกติที่ตรวจจับได้ของข้อมูลในระบบที่มีสถานะสำเร็จ และล้มเหลว จะพบว่าในแผนภูมิรูปวงกลมวิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง วิธีไอโซเลชั่นฟอเรส และวิธีไอเอฟ-แอลโอเอฟ ให้สัดส่วนรายการสิ่งผิดปรกติใกล้เคียงกัน ยกเว้นวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ที่ให้สัดส่วนแตกต่างกัน ซึ่งสถานะสำเร็จให้สัดส่วนร้อยละ 1.53 และสถานะล้มเหลวให้สัดส่วนร้อยละ 1.97

ตารางที่ 4.4 วิธีการตรวจจับสิ่งผิดปรกติที่ให้ผลลัพธ์เหมือนกันของการเข้ามาใช้งานในระบบ

ชุดข้อมูล	วิธีการตรวจจับสิ่งผิดปรกติ				ผลลัพธ์ที่เหมือนกัน
รายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จ	LOF	OCSVM	Isolation-Forest	IF-LOF	177
รายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลว	LOF	OCSVM	Isolation-Forest	IF-LOF	162

จากตารางที่ 4.4 ผลลัพธ์ที่ตรวจจับได้ตรงกันโดยทั้ง 4 วิธี อาจถูกสงสัยว่าเป็นสิ่งผิดปรกติ ที่ควรได้รับการพิจารณาและตรวจสอบในเบื้องต้นก่อน ซึ่งเป็นการช่วยลดขั้นตอนการตรวจสอบสิ่งผิดปรกติของบริษัท

4.2.1 สถิติพรรณนาสิ่งผิดปรกติของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จ

4.2.1.1 วิธีโลคอลเอาท์ไลเออร์แฟคเตอร์

ตารางที่ 4.5 สถิติพรรณนาสิ่งผิดปรกติของวิธี LOF รายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จ

สถิติพรรณนาสิ่งผิดปรกติของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จ วิธี LOF จำนวน 1,771 รายการ					
ตัวแปร	ค่าเฉลี่ย	SD	Median	IQR	Skewness
ช่วงเวลาเช้า	18.76	80.46	4	13	11.98
ช่วงเวลากลางวัน	17.80	90.87	5	13	15.25
ช่วงเวลาเย็น	7.57	66.97	1	3	17.88
ช่วงเวลาดึก	3.93	39.64	0	1	28.74

4.2.1.2 วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง

ตารางที่ 4.6 สถิติพรรณนาสิ่งผิดปรกติของวิธี OCSVM รายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จ

สถิติพรรณนาสิ่งผิดปรกติของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จ วิธี OCSVM จำนวน 2,160 รายการ					
ตัวแปร	ค่าเฉลี่ย	SD	Median	IQR	Skewness
ช่วงเวลาเช้า	39.57	78.26	25	52.75	10.12
ช่วงเวลากลางวัน	40.02	88.24	25	54	13.10
ช่วงเวลาเย็น	14.53	64.53	0	15	16.29
ช่วงเวลาดึก	4.47	36.11	0	0	31.12

4.2.1.3 วิธีไอโซเลชั่นฟอเรส

ตารางที่ 4.7 สถิติพรรณนาสิ่งผิดปรกติของวิธี Isolation Forest รายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จ

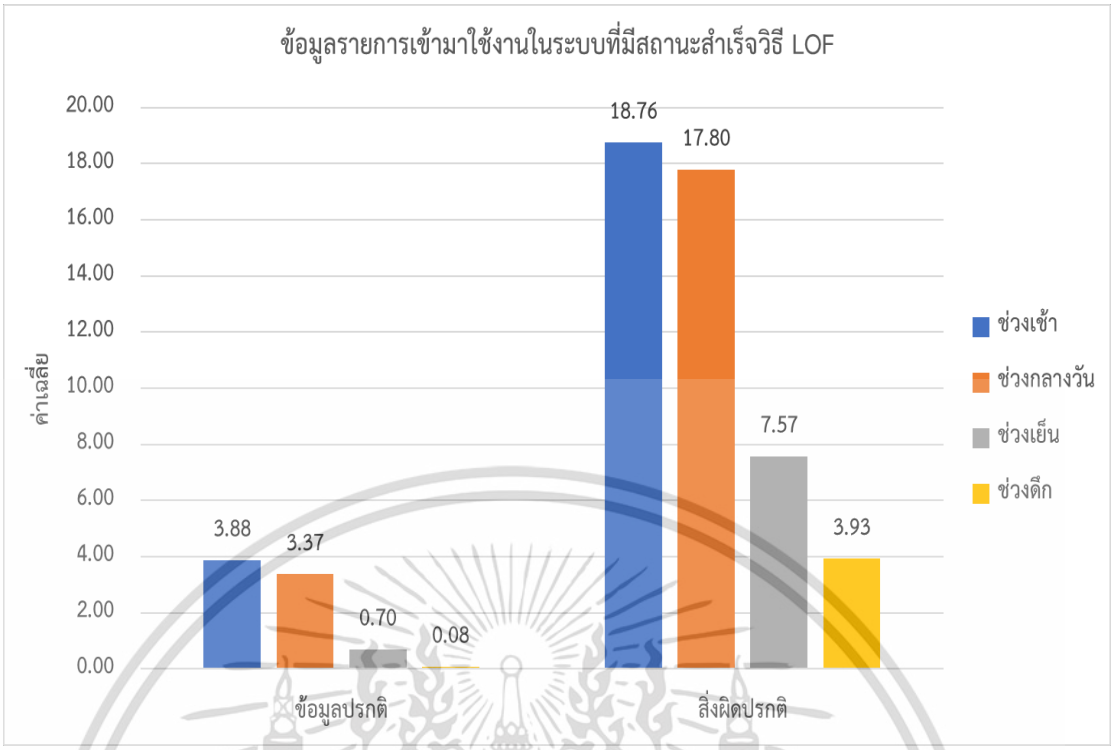
สถิติพรรณนาสิ่งผิดปรกติของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จ วิธี Isolation Forest จำนวน 2,315 รายการ					
ตัวแปร	ค่าเฉลี่ย	SD	Median	IQR	Skewness
ช่วงเวลาเช้า	33.78	75.90	15	39	10.57
ช่วงเวลากลางวัน	37.18	85.40	19	45	13.58
ช่วงเวลาเย็น	15.13	62.20	3	15	16.94
ช่วงเวลาดึก	4.92	34.86	0	5	32.22

4.2.1.4 วิธีไอเอฟ-แอลโอเอฟ

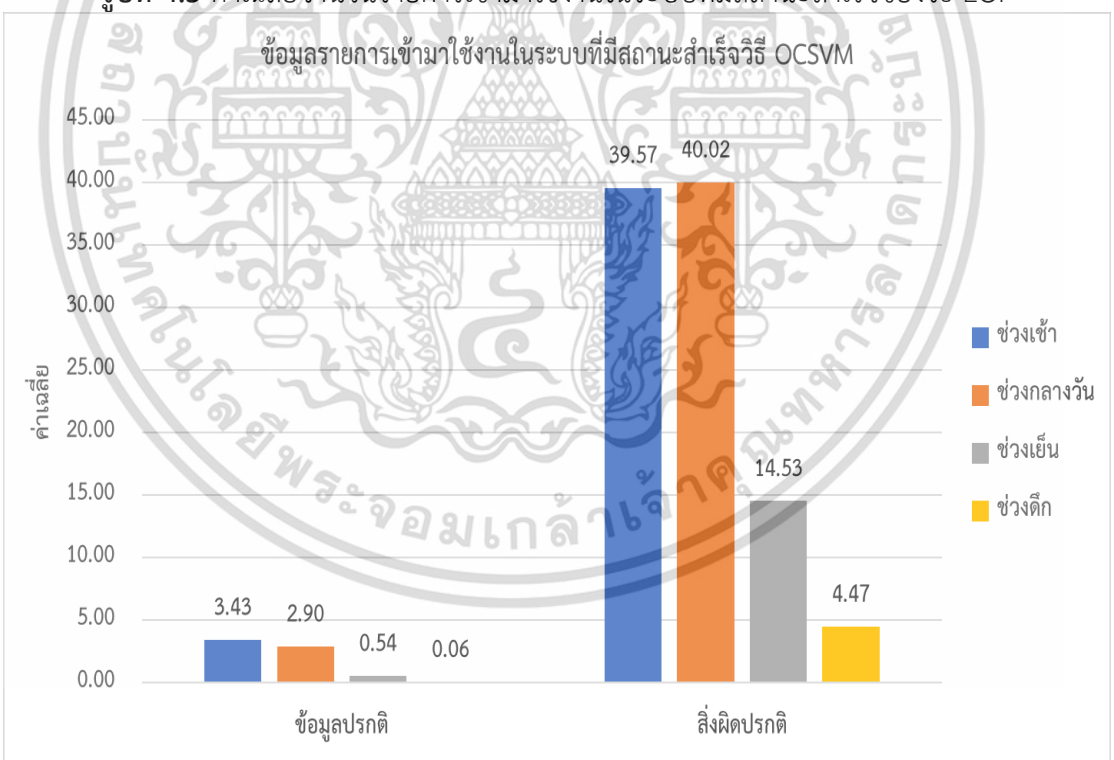
ตารางที่ 4.8 สถิติพรรณนาสิ่งผิดปรกติของวิธี IF-LOF รายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จ

สถิติพรรณนาสิ่งผิดปรกติของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จ วิธี IF-LOF จำนวน 228 รายการ					
ตัวแปร	ค่าเฉลี่ย	SD	Median	IQR	Skewness
ช่วงเวลาเช้า	89.28	209.89	7.5	81	4.25
ช่วงเวลากลางวัน	81.04	243.91	7	50.75	5.39
ช่วงเวลาเย็น	45.55	182.58	2	22.75	6.30
ช่วงเวลาดึก	27.93	107.64	6	12.75	10.59

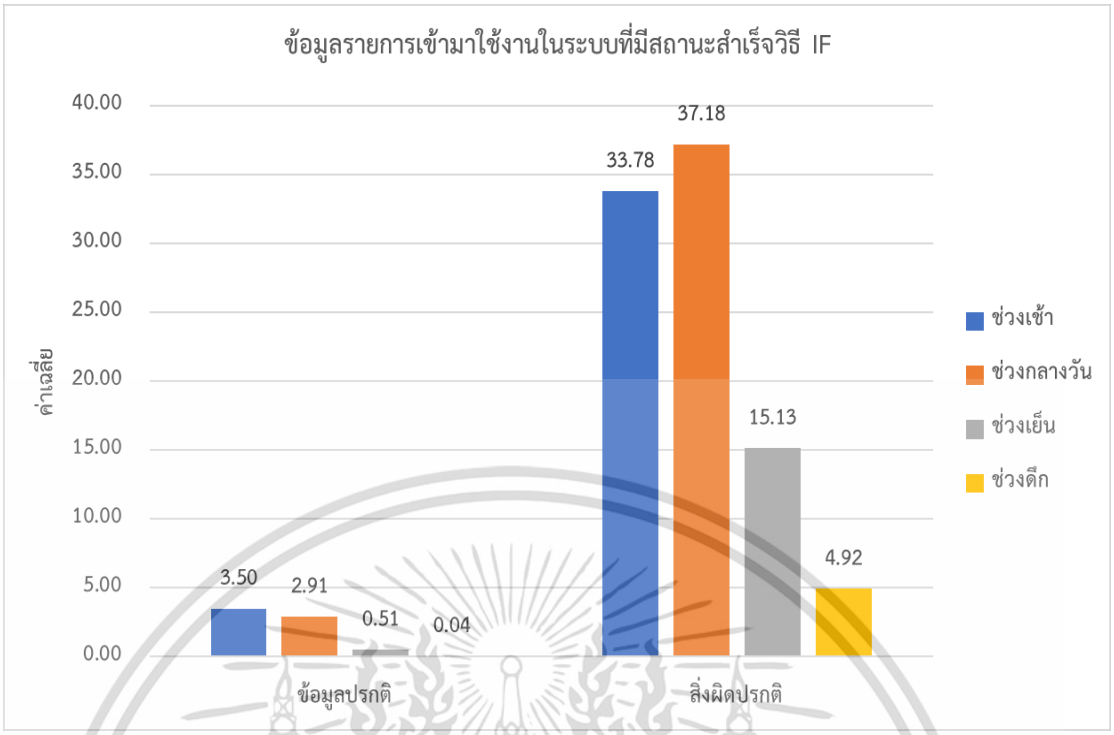
จากตารางที่ 4.5 - 4.8 จะพบว่าค่าสิ่งผิดปรกติโดยเฉลี่ยสำหรับรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จด้วยวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง วิธีไอโซเลชั่นฟอเรส และวิธีไอเอฟ-แอลโอเอฟ ในช่วงเวลาเช้า ช่วงเวลากลางวันสูงกว่าช่วงเวลาเย็นและช่วงเวลาดึก นอกจากนี้ยังพบว่าส่วนเบี่ยงเบนมาตรฐานในแต่ละช่วงเวลามีค่าสูง แสดงให้เห็นว่าค่าของสิ่งผิดปรกติมีการกระจายตัวมาก โดยเฉพาะวิธีไอเอฟ-แอลโอเอฟ มีจำนวนรายการสิ่งผิดปรกติของการเข้ามาใช้งานในระบบที่ตรวจจับได้มีค่าส่วนเบี่ยงเบนมาตรฐานสูงกว่าทั้ง 3 วิธีในทุกช่วงเวลา เนื่องจากวิธีไอเอฟ-แอลโอเอฟจะนำสิ่งผิดปรกติของวิธีไอโซเลชั่นฟอเรสมาตรวจจับต่อ ทำให้ข้อมูลที่ได้มีจำนวนรายการของสิ่งผิดปรกติที่น้อยลง ซึ่งส่งผลให้ข้อมูลมีการกระจายตัวมาก และทั้ง 4 วิธีมีลักษณะเบ้ขวาที่สามารถดูได้จากค่า Skewness ที่คำนวณได้มีค่าเป็นบวก นั่นคือ ค่าสิ่งผิดปรกติที่มีค่าน้อยมีความถี่มากกว่าค่าสิ่งผิดปรกติที่มีค่ามาก โดยเมื่อนำไปสร้างแผนภูมิแท่งจะมีลักษณะดังรูปที่ 4.3 - 4.6



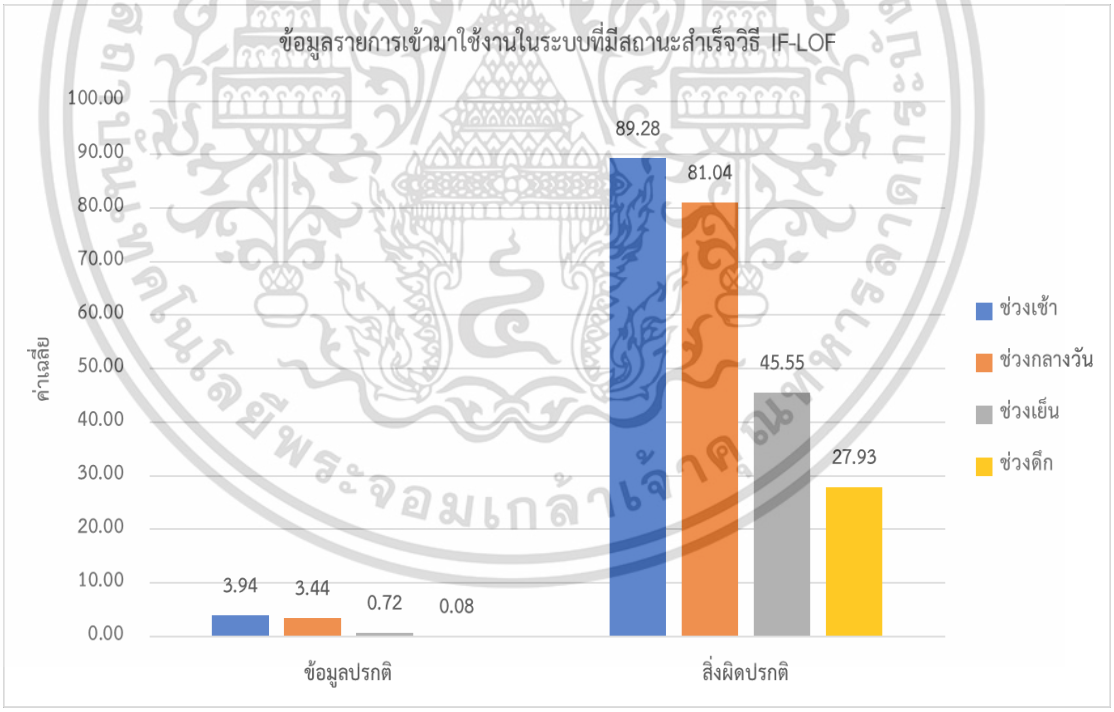
รูปที่ 4.3 ค่าเฉลี่ยจำนวนรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จของวิธี LOF



รูปที่ 4.4 ค่าเฉลี่ยจำนวนรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จของวิธี OCSVM



รูปที่ 4.5 ค่าเฉลี่ยจำนวนรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จของวิธี Isolation Forest



รูปที่ 4.6 ค่าเฉลี่ยจำนวนรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จของวิธี IF-LOF

จากรูปที่ 4.3 – 4.6 จะเห็นได้ว่ามีค่าเฉลี่ยของสิ่งผิดปรกติที่ตรวจจับโดยวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง วิธีไอโซเลชั่นฟอเรส และวิธีไอเอฟ-แอลโอเอฟ สูงกว่าข้อมูลปรกติเป็นอย่างมากในทุกช่วงเวลา โดยเฉพาะช่วงเวลาเช้าและกลางวัน

4.2.2 สถิติพรรณนาสิ่งผิดปรกติของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลว

4.2.2.1 วิธีโลคอลเอาท์ไลเออร์แฟคเตอร์

ตารางที่ 4.9 สถิติพรรณนาสิ่งผิดปรกติของวิธี LOF รายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลว

สถิติพรรณนาสิ่งผิดปรกติของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลว วิธี LOF จำนวน 1,070 รายการ					
ตัวแปร	ค่าเฉลี่ย	SD	Median	IQR	Skewness
ช่วงเวลาเช้า	14.97	52.89	2	8	7.00
ช่วงเวลากลางวัน	14.48	49.84	1	7	6.82
ช่วงเวลาเย็น	10.60	46.57	1	4	7.98
ช่วงเวลาดึก	8.84	49.60	0	1	9.28

4.2.2.2 วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง

ตารางที่ 4.10 สถิติพรรณนาสิ่งผิดปรกติของวิธี OCSVM รายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลว

สถิติพรรณนาสิ่งผิดปรกติของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลว วิธี OCSVM จำนวน 1,003 รายการ					
ตัวแปร	ค่าเฉลี่ย	SD	Median	IQR	Skewness
ช่วงเวลาเช้า	31.87	64.26	9	37	4.53
ช่วงเวลากลางวัน	28.11	61.56	7	30	4.89
ช่วงเวลาเย็น	20.28	56.94	0	18	5.51
ช่วงเวลาดึก	14.05	56.07	0	8	7.30

4.2.2.3 วิธีไอโซเลชั่นฟอเรส

ตารางที่ 4.11 สถิติพรรณนาสิ่งผิดปรกติของวิธี Isolation Forest รายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลว

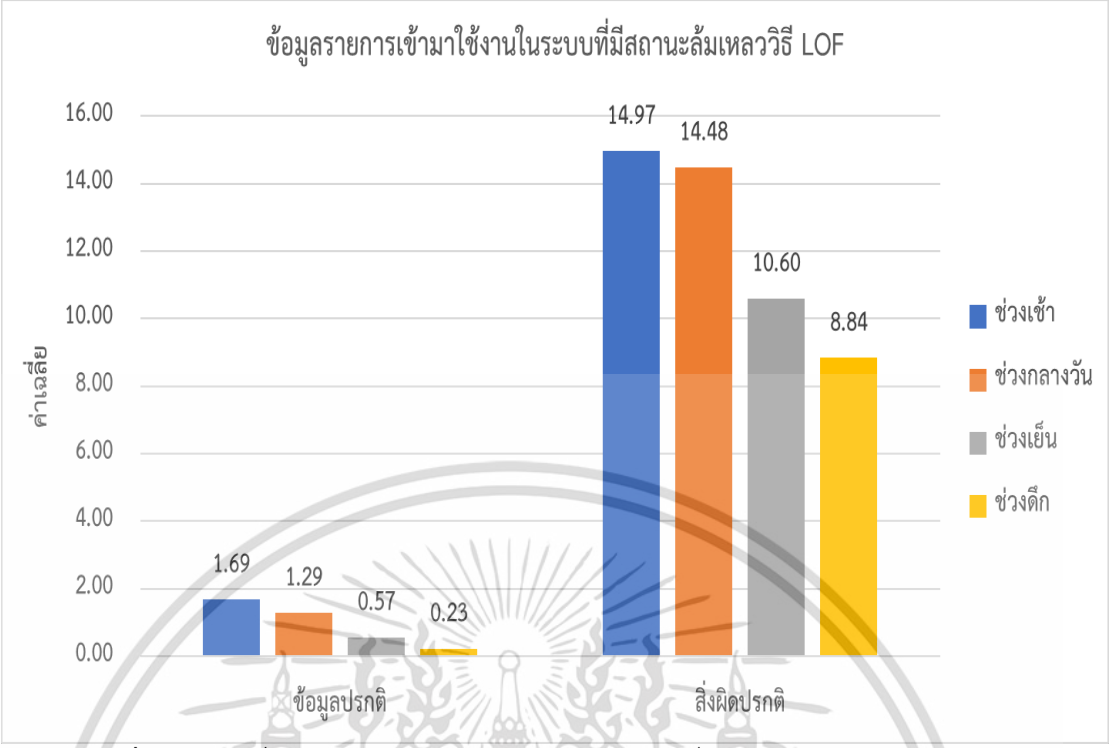
สถิติพรรณนาสิ่งผิดปรกติของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลว วิธี Isolation Forest จำนวน 1,081 รายการ					
ตัวแปร	ค่าเฉลี่ย	SD	Median	IQR	Skewness
ช่วงเวลาเช้า	27.40	62.24	8	19	4.84
ช่วงเวลากลางวัน	26.22	59.35	8	22	5.16
ช่วงเวลาเย็น	19.93	54.69	6	13	5.79
ช่วงเวลาดึก	14.50	53.84	2	8.5	7.62

4.2.2.4 วิธีไอเอฟ-แอลโอเอฟ

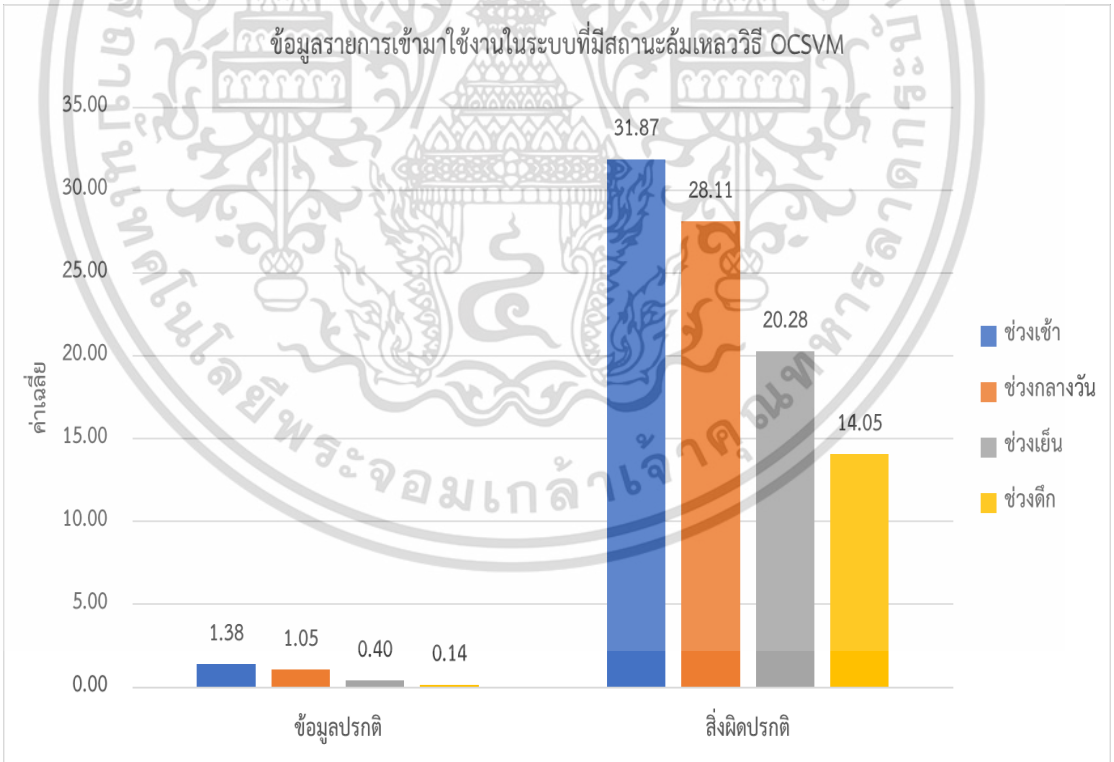
ตารางที่ 4.12 สถิติพรรณนาสิ่งผิดปรกติของวิธี IF-LOF รายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลว

สถิติพรรณนาสิ่งผิดปรกติของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลว วิธี IF-LOF จำนวน 167 รายการ					
ตัวแปร	ค่าเฉลี่ย	SD	Median	IQR	Skewness
ช่วงเวลาเช้า	71.04	118.13	20	88	2.54
ช่วงเวลากลางวัน	69.51	110.05	21	107	2.48
ช่วงเวลาเย็น	55.27	107.08	12	63	2.93
ช่วงเวลาดึก	51.28	117.04	11	38	3.45

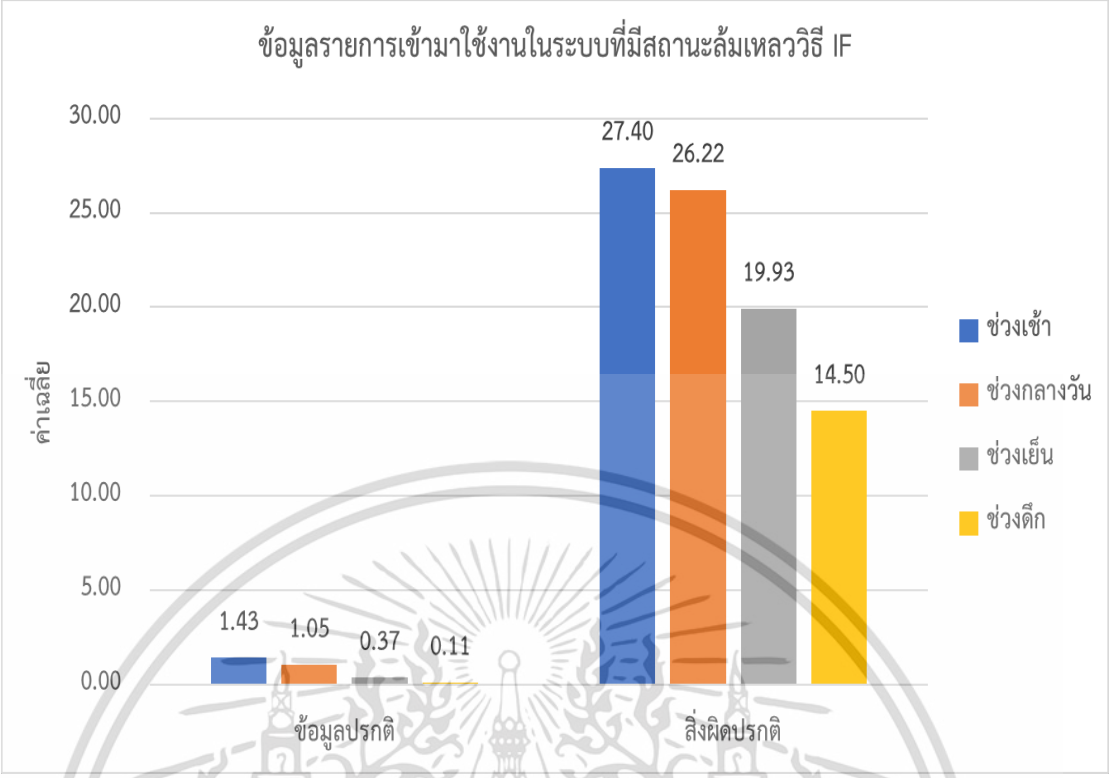
จากตารางที่ 4.9 - 4.12 จะพบว่าค่าสิ่งผิดปรกติโดยเฉลี่ยสำหรับรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลวด้วยวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่งวิธีไอโซเลชั่นฟอเรส และวิธีไอเอฟ-แอลโอเอฟ ในช่วงเวลาเช้า กลางวัน และเย็น สูงกว่าช่วงเวลาดึก นอกจากนี้ยังพบว่าส่วนเบี่ยงเบนมาตรฐานในแต่ละช่วงเวลามีค่าสูง แสดงให้เห็นว่าค่าของสิ่งผิดปรกติมีการกระจายตัวมาก โดยเฉพาะวิธีไอเอฟ-แอลโอเอฟ มีจำนวนรายการสิ่งผิดปรกติของการเข้ามาใช้งานในระบบที่ตรวจจับได้มีค่าส่วนเบี่ยงเบนมาตรฐานสูงกว่าทั้ง 3 วิธีในทุกช่วงเวลา เนื่องจากวิธีไอเอฟ-แอลโอเอฟจะนำสิ่งผิดปรกติของวิธีไอโซเลชั่นฟอเรสมาตรวจจับต่อ ทำให้ข้อมูลที่ได้มีจำนวนรายการของสิ่งผิดปรกติที่น้อยลง ซึ่งส่งผลให้ข้อมูลมีการกระจายตัวมาก และทั้ง 4 วิธีมีลักษณะเบ้ขวาที่สามารถดูได้จากค่า Skewness ที่คำนวณได้มีค่าเป็นบวก นั่นคือ ค่าสิ่งผิดปรกติที่มีค่าน้อยมีความถี่มากกว่าค่าสิ่งผิดปรกติที่มีค่ามาก โดยเมื่อนำไปสร้างแผนภูมิแท่งจะมีลักษณะดังรูปที่ 4.7 - 4.10



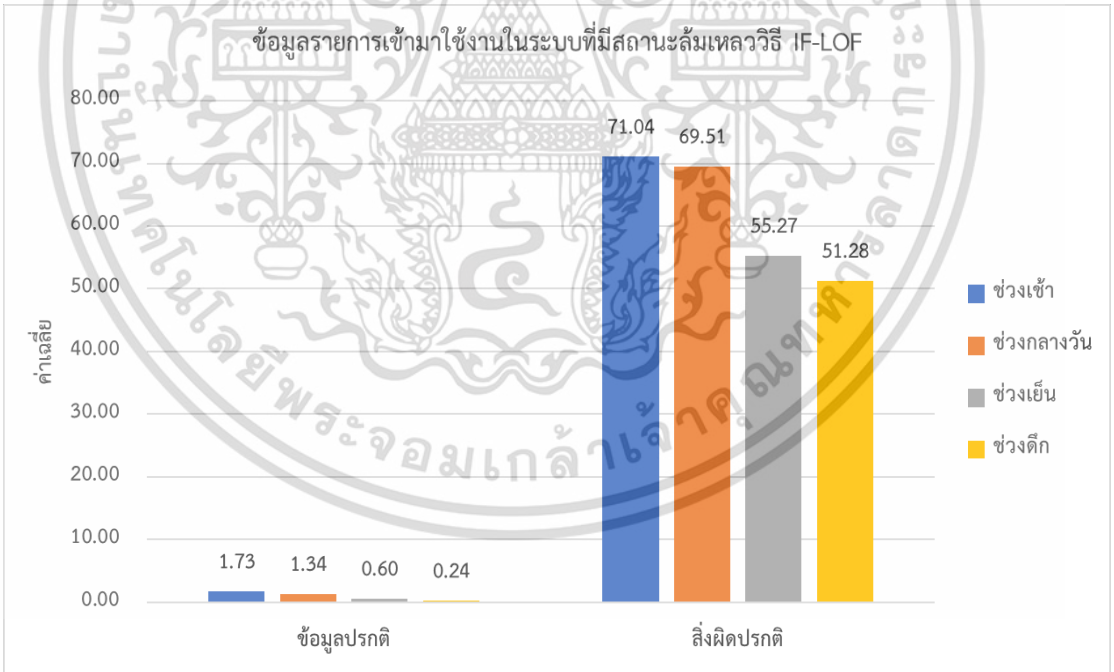
รูปที่ 4.7 ค่าเฉลี่ยจำนวนรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลวของวิธี LOF



รูปที่ 4.8 ค่าเฉลี่ยจำนวนรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลวของวิธี OCSVM



รูปที่ 4.9 ค่าเฉลี่ยจำนวนรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลวของวิธี Isolation Forest



รูปที่ 4.10 ค่าเฉลี่ยจำนวนรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลวของวิธี IF-LOF

จากรูปที่ 4.7 – 4.10 จะเห็นได้ว่ามีค่าเฉลี่ยของสิ่งผิดปรกติที่ตรวจจับ โดยวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง วิธีไอโซเลชั่นฟอเรส และวิธีไอเอฟ-แอลโอเอฟ สูงกว่าข้อมูลปรกติเป็นอย่างมากในทุกช่วงเวลา โดยเฉพาะช่วงเวลาเช้าและช่วงเวลากลางวัน

4.3. ผลการวิเคราะห์การจัดกลุ่ม (Cluster Analysis)

หลังจากที่ได้สิ่งผิดปรกติจากวิธีการตรวจจับสิ่งผิดปรกติทั้ง 4 วิธี แล้วนั้นนำไปวิเคราะห์ผลเพื่อดูคุณลักษณะของสิ่งผิดปรกติ โดยอาศัยเทคนิคการจัดกลุ่มข้อมูลด้วยวิธีเคมีน โดยวิธีนี้จะแบ่งข้อมูลออกเป็นกลุ่มย่อย ๆ ที่มีลักษณะคล้ายกัน โดยผู้วิจัยจะมีการกำหนดค่า K หรือจำนวนกลุ่มด้วยวิธี Elbow วัดจากค่าความคลาดเคลื่อนของผลรวมระยะห่างระหว่างวัตถุกับจุดศูนย์กลางเริ่มต้น หรือค่าผลบวกกำลังสองภายในกลุ่ม และวิธี Silhouette โดยใช้ค่าเฉลี่ยของระยะห่างระหว่างจุดกับจุดต่าง ๆ ภายในกลุ่มเดียวกัน ส่วนด้วยระยะห่างน้อยที่สุดของจุดกับจุดต่าง ๆ ในแต่ละกลุ่ม เพื่อที่จะหาจำนวนกลุ่ม (K) ที่เหมาะสมที่สุด ซึ่งทางผู้วิจัยได้นำเสนอผลลัพธ์การจัดกลุ่มของแต่ละแบบจำลองในข้อมูล 2 ชุดดังนี้

4.3.1 ข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จ

4.3.1.1 วิธีโลคอลเอาท์ไลเออร์แฟคเตอร์

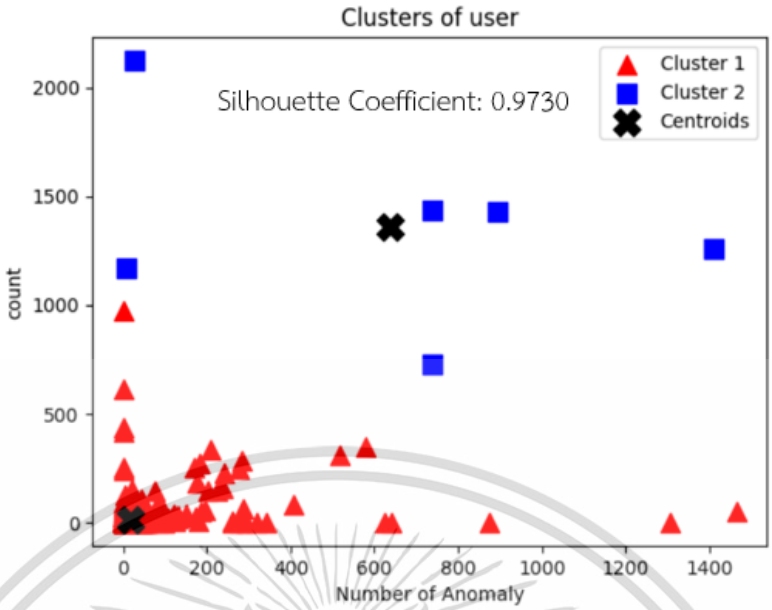
ตารางที่ 4.13 ผลการจัดกลุ่มของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จของวิธี LOF

ผลการจัดกลุ่มของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จของวิธี LOF				
ตัวแปร	Cluster1 n=1,765		Cluster2 n=6	
	ค่าเฉลี่ย	SD	ค่าเฉลี่ย	SD
ช่วงเวลาเช้า	16.65	66.13	635.83	539.14
ช่วงเวลากลางวัน	13.23	39.57	1,358.00	456.35
ช่วงเวลาเย็น	4.67	31.56	861.00	599.64
ช่วงเวลาดึก	3.89	39.69	12.33	23.63

จากตารางที่ 4.13 แสดงค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐาน ของแต่ละช่วงเวลากำหนดไว้ 2 กลุ่ม (Cluster) ผลการจัดกลุ่มสรุปได้ดังนี้

กลุ่มที่ 1 คือ กลุ่มของสิ่งผิดปรกติที่มีพฤติกรรมการเข้ามาใช้งานเฉลี่ยน้อยกว่าค่าเฉลี่ยของสิ่งผิดปรกติที่ตรวจจับได้ทั้งหมด และค่าส่วนเบี่ยงเบนมาตรฐานที่น้อยกว่า ซึ่งแสดงว่ากลุ่มที่ 1 ข้อมูลมีการเกาะกลุ่มกัน

กลุ่มที่ 2 คือ กลุ่มของสิ่งผิดปรกติที่มีพฤติกรรมการเข้ามาใช้งานเฉลี่ยเป็นอย่างมากเมื่อเทียบกับค่าเฉลี่ยของสิ่งผิดปรกติที่ตรวจจับได้ทั้งหมด และค่าส่วนเบี่ยงเบนมาตรฐานในช่วงเวลาเช้า กลางวัน เย็นที่มากกว่าซึ่งแสดงว่าช่วงเวลาเช้า กลางวัน เย็นของข้อมูลในกลุ่มที่ 2 มีการกระจายตัวมาก แต่ในส่วนของช่วงเวลาดึกนั้นมีการกระจายตัวน้อย



รูปที่ 4.11 การแบ่งกลุ่มของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จของวิธี LOF

จากรูปที่ 4.11 จะพบว่าข้อมูลส่วนใหญ่ของกลุ่มที่ 1 ข้อมูลมีการเกาะกลุ่มกันและมีการกระจายตัวที่น้อย และกลุ่มที่ 2 ข้อมูลมีการกระจายตัวที่มาก และค่า Silhouette Coefficient มีค่าเท่ากับ 0.973 ซึ่งมีค่าใกล้ 1 แสดงว่าข้อมูลถูกจัดให้อยู่ในกลุ่มมีความเหมาะสมแล้ว ในส่วนของรูปการแบ่งกลุ่มของข้อมูลรายการเข้ามาใช้งานในระบบสามารถดูการแบ่งกลุ่มของวิธีการทั้งหมดที่ภาคผนวก ง

4.3.1.2 วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง

ตารางที่ 4.14 ผลการจัดกลุ่มของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จของวิธี OCSVM

ผลการจัดกลุ่มของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จของวิธี OCSVM				
ตัวแปร	Cluster1 n=2,154		Cluster2 n=6	
	ค่าเฉลี่ย	SD	ค่าเฉลี่ย	SD
ช่วงเวลาเช้า	37.91	66.88	635.83	539.14
ช่วงเวลากลางวัน	36.34	49.71	1,358.00	456.35
ช่วงเวลาเย็น	12.16	36.58	861.00	599.64
ช่วงเวลาดึก	4.45	36.14	12.33	23.63

จากตารางที่ 4.14 แสดงค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐาน ของแต่ละช่วงเวลากำหนดไว้ 2 กลุ่ม ผลการจัดกลุ่มสรุปได้ดังนี้

กลุ่มที่ 1 คือ กลุ่มของสิ่งผิดปรกติที่มีพฤติกรรมการเข้ามาใช้งานเฉลี่ยน้อยกว่าค่าเฉลี่ยของสิ่งผิดปรกติที่ตรวจจับได้ทั้งหมด และค่าส่วนเบี่ยงเบนมาตรฐานที่น้อยกว่า ซึ่งแสดงว่ากลุ่มที่ 1 ข้อมูลมีการเกาะกลุ่มกัน

กลุ่มที่ 2 คือ กลุ่มของสิ่งผิดปรกติที่มีพฤติกรรมการเข้ามาใช้งานเฉลี่ยเป็นอย่างมากเมื่อเทียบกับค่าเฉลี่ยของสิ่งผิดปรกติที่ตรวจจับได้ทั้งหมด และค่าส่วนเบี่ยงเบนมาตรฐานในช่วงเวลาเช้า กลางวัน เย็นที่มากกว่าซึ่งแสดงว่าช่วงเวลาเช้า กลางวัน เย็นของข้อมูลในกลุ่มที่ 2 มีการกระจายตัวมาก แต่ในส่วนของช่วงเวลาดึกนั้นมีการกระจายตัวน้อย

4.3.1.3 วิธีไอโซเลชั่นฟอเรส

ตารางที่ 4.15 ผลการจัดกลุ่มของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จของวิธี Isolation Forest

ผลการจัดกลุ่มของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จของวิธี Isolation Forest				
ตัวแปร	Cluster1 n=2,309		Cluster2 n=6	
	ค่าเฉลี่ย	SD	ค่าเฉลี่ย	SD
ช่วงเวลาเช้า	32.21	64.82	635.83	539.14
ช่วงเวลากลางวัน	33.74	48.11	1,358.00	456.35
ช่วงเวลาเย็น	12.92	35.15	861.00	599.64
ช่วงเวลาดึก	4.90	34.89	12.33	23.63

จากตารางที่ 4.15 แสดงค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐาน ของแต่ละช่วงเวลากำหนดไว้ 2 กลุ่ม ผลการจัดกลุ่มสรุปได้ดังนี้

กลุ่มที่ 1 คือ กลุ่มของสิ่งผิดปรกติที่มีพฤติกรรมการเข้ามาใช้งานเฉลี่ยน้อยกว่าค่าเฉลี่ยของสิ่งผิดปรกติที่ตรวจจับได้ทั้งหมด และค่าส่วนเบี่ยงเบนมาตรฐานที่น้อยกว่า ซึ่งแสดงว่ากลุ่มที่ 1 ข้อมูลมีการเกาะกลุ่มกัน

กลุ่มที่ 2 คือ กลุ่มของสิ่งผิดปรกติที่มีพฤติกรรมการเข้ามาใช้งานเฉลี่ยเป็นอย่างมากเมื่อเทียบกับค่าเฉลี่ยของสิ่งผิดปรกติที่ตรวจจับได้ทั้งหมด และค่าส่วนเบี่ยงเบนมาตรฐานในช่วงเวลาเช้า กลางวัน เย็นที่มากกว่าซึ่งแสดงว่าช่วงเวลาเช้า กลางวัน เย็นของข้อมูลในกลุ่มที่ 2 มีการกระจายตัวมาก แต่ในส่วนของช่วงเวลาดึกนั้นมีการกระจายตัวน้อย

4.3.1.4 วิธีไอเอฟ-แอลโอเอฟ

ตารางที่ 4.16 ผลการจัดกลุ่มของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จของวิธี IF-LOF

ผลการจัดกลุ่มของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จของวิธี IF-LOF				
ตัวแปร	Cluster1 n=222		Cluster2 n=6	
	ค่าเฉลี่ย	SD	ค่าเฉลี่ย	SD
ช่วงเวลาเช้า	74.50	174.20	635.83	539.14
ช่วงเวลากลางวัน	46.52	104.55	1,358.00	456.35
ช่วงเวลาเย็น	23.50	86.97	861.00	599.64
ช่วงเวลาดึก	28.34	109.00	12.33	23.63

จากตารางที่ 4.16 แสดงค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐาน ของแต่ละช่วงเวลากำหนดไว้ 2 กลุ่ม ผลการจัดกลุ่มสรุปได้ดังนี้

กลุ่มที่ 1 คือ กลุ่มของสิ่งผิดปรกติที่มีพฤติกรรมการเข้ามาใช้งานเฉลี่ยน้อยกว่าค่าเฉลี่ยของสิ่งผิดปรกติที่ตรวจจับได้ทั้งหมด และค่าส่วนเบี่ยงเบนมาตรฐานที่น้อยกว่า ซึ่งแสดงว่ากลุ่มที่ 1 ข้อมูลมีการเกาะกลุ่มกัน

กลุ่มที่ 2 คือ กลุ่มของสิ่งผิดปรกติที่มีพฤติกรรมการเข้ามาใช้งานเฉลี่ยเป็นอย่างมากเมื่อเทียบกับค่าเฉลี่ยของสิ่งผิดปรกติที่ตรวจจับได้ทั้งหมด และค่าส่วนเบี่ยงเบนมาตรฐานในช่วงเวลาเช้า กลางวัน เย็นที่มากกว่าซึ่งแสดงว่าช่วงเวลาเช้า กลางวัน เย็นของข้อมูลในกลุ่มที่ 2 มีการกระจายตัวมาก แต่ในส่วนของช่วงเวลาดึกนั้นมีการกระจายตัวน้อย

4.3.2 ข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลว

4.3.2.1 วิธีโลคอลเอาท์ไลเออร์แฟคเตอร์

ตารางที่ 4.17 ผลการจัดกลุ่มของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลวของวิธี LOF

ผลการจัดกลุ่มของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลวของวิธี LOF				
ลักษณะปัจจัย	Cluster1 n=1,064		Cluster2 n=6	
	ค่าเฉลี่ย	SD	ค่าเฉลี่ย	SD
ช่วงเวลาเช้า	12.42	40.62	465.16	51.64
ช่วงเวลากลางวัน	11.78	34.42	493.83	44.61
ช่วงเวลาเย็น	7.86	28.66	496.66	62.65
ช่วงเวลาดึก	5.89	30.11	531.50	59.84

จากตารางที่ 4.17 แสดงค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐาน ของแต่ละช่วงเวลากำหนดไว้ 2 กลุ่ม ผลการจัดกลุ่มสรุปได้ดังนี้

กลุ่มที่ 1 คือ กลุ่มของสิ่งผิดปรกติที่มีพฤติกรรมการเข้ามาใช้งานเฉลี่ยน้อยกว่าค่าเฉลี่ยของสิ่งผิดปรกติที่ตรวจจับได้ทั้งหมด

กลุ่มที่ 2 คือ กลุ่มของสิ่งผิดปรกติที่มีพฤติกรรมการเข้ามาใช้งานเฉลี่ยเป็นอย่างมากเมื่อเทียบกับค่าเฉลี่ยของสิ่งผิดปรกติที่ตรวจจับได้ทั้งหมด

4.3.2.2 วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง

ตารางที่ 4.18 ผลการจัดกลุ่มของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลวของวิธี OCSVM

ผลการจัดกลุ่มของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลวของวิธี OCSVM				
ลักษณะปัจจัย	Cluster1 n=974		Cluster2 n=29	
	ค่าเฉลี่ย	SD	ค่าเฉลี่ย	SD
ช่วงเวลาเช้า	23.33	35.64	318.55	126.46
ช่วงเวลากลางวัน	20.28	32.91	291.00	155.83
ช่วงเวลาเย็น	13.68	30.90	241.79	174.70
ช่วงเวลาดึก	6.91	18.83	253.79	197.25

จากตารางที่ 4.18 แสดงค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐาน ของแต่ละช่วงเวลากำหนดไว้ 2 กลุ่ม ผลการจัดกลุ่มสรุปได้ดังนี้

กลุ่มที่ 1 คือ กลุ่มของสิ่งผิดปรกติที่มีพฤติกรรมการเข้ามาใช้งานเฉลี่ยน้อยกว่าค่าเฉลี่ยของสิ่งผิดปรกติที่ตรวจจับได้ทั้งหมด และค่าส่วนเบี่ยงเบนมาตรฐานที่น้อยกว่า ซึ่งแสดงว่ากลุ่มที่ 1 ข้อมูลมีการเกาะกลุ่มกัน

กลุ่มที่ 2 คือ กลุ่มของสิ่งผิดปรกติที่มีพฤติกรรมการเข้ามาใช้งานเฉลี่ยที่มากกว่าเมื่อเทียบกับค่าเฉลี่ยของสิ่งผิดปรกติที่ตรวจจับได้ทั้งหมด และค่าส่วนเบี่ยงเบนมาตรฐานที่มากกว่าซึ่งแสดงว่ากลุ่มที่ 2 ข้อมูลมีการกระจายตัวกัน

4.3.2.3 วิธีไอโซเลชั่นฟอเรส

ตารางที่ 4.19 ผลการจัดกลุ่มของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลวของวิธี Isolation Forest

ผลการจัดกลุ่มของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลวของวิธี Isolation Forest				
ลักษณะปัจจัย	Cluster1 n=1,052		Cluster2 n=29	
	ค่าเฉลี่ย	SD	ค่าเฉลี่ย	SD
ช่วงเวลาเช้า	19.36	33.93	318.55	126.46
ช่วงเวลากลางวัน	18.92	31.38	291.00	155.83
ช่วงเวลาเย็น	13.81	29.41	241.79	174.70
ช่วงเวลาดึก	7.90	17.87	253.79	197.25

จากตารางที่ 4.19 แสดงค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐาน ของแต่ละช่วงเวลากำหนดไว้ 2 กลุ่ม ผลการจัดกลุ่มสรุปได้ดังนี้

กลุ่มที่ 1 คือ กลุ่มของสิ่งผิดปรกติที่มีพฤติกรรมการเข้ามาใช้งานเฉลี่ยน้อยกว่าค่าเฉลี่ยของสิ่งผิดปรกติที่ตรวจจับได้ทั้งหมด และค่าส่วนเบี่ยงเบนมาตรฐานที่น้อยกว่า ซึ่งแสดงว่ากลุ่มที่ 1 ข้อมูลมีการเกาะกลุ่มกัน

กลุ่มที่ 2 คือ กลุ่มของสิ่งผิดปรกติที่มีพฤติกรรมการเข้ามาใช้งานเฉลี่ยที่มากกว่าเมื่อเทียบกับค่าเฉลี่ยของสิ่งผิดปรกติที่ตรวจจับได้ทั้งหมด และค่าส่วนเบี่ยงเบนมาตรฐานที่มากกว่า ซึ่งแสดงว่ากลุ่มที่ 2 ข้อมูลมีการกระจายตัวกัน

4.3.2.4 วิธีไอเอฟ-แอลโอเอฟ

ตารางที่ 4.20 ผลการจัดกลุ่มของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลวของวิธี IF-LOF

ผลการจัดกลุ่มของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลวของวิธี IF-LOF				
ลักษณะปัจจัย	Cluster1 n=161		Cluster2 n=6	
	ค่าเฉลี่ย	SD	ค่าเฉลี่ย	SD
ช่วงเวลาเช้า	56.35	91.39	465.16	51.64
ช่วงเวลากลางวัน	53.70	74.16	493.83	44.61
ช่วงเวลาเย็น	38.81	64.78	496.66	62.65
ช่วงเวลาดึก	33.37	71.62	531.50	59.84

จากตารางที่ 4.20 แสดงค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐาน ของแต่ละช่วงเวลากำหนดไว้ 2 กลุ่ม ผลการจัดกลุ่มสรุปได้ดังนี้

กลุ่มที่ 1 คือ กลุ่มของสิ่งผิดปรกติที่มีพฤติกรรมการเข้ามาใช้งานเฉลี่ยน้อยกว่าค่าเฉลี่ยของสิ่งผิดปรกติที่ตรวจจับได้ทั้งหมด

กลุ่มที่ 2 คือ กลุ่มของสิ่งผิดปรกติที่มีพฤติกรรมการเข้ามาใช้งานเฉลี่ยเป็นอย่างมากเมื่อเทียบกับค่าเฉลี่ยของสิ่งผิดปรกติที่ตรวจจับได้ทั้งหมด

4.4. ผลการวิเคราะห์ความแตกต่างระหว่างกลุ่มโดยสถิติอนุมาน (Inferential Statistics)

4.4.1 การทดสอบสมมติฐานจากการจัดกลุ่มข้อมูลการเข้ามาใช้งานระบบที่มีสถานะสำเร็จ

4.4.1.1 การทดสอบสมมติฐานจากการจัดกลุ่มของวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์

การวิเคราะห์ทางสถิติเพื่อตรวจว่าคุณลักษณะหรือตัวแปรสามารถแบ่งค่าความแตกต่างระหว่างกลุ่มได้ ของการเข้ามาใช้งานระบบในแต่ละช่วงเวลา โดยใช้สถิติทดสอบ Mann-Whitney U ด้วยโปรแกรมสำเร็จรูป IBM SPSS ได้ผลลัพธ์ดังต่อไปนี้

สมมติฐานการทดสอบ

$$H_0 : M_1 = M_2$$

$$H_1 : M_1 \neq M_2$$

ตารางที่ 4.21 ผลการทดสอบความแตกต่างระหว่างกลุ่มของวิธี LOF

ตัวแปร	กลุ่ม (Cluster)	n	Mann-Whitney U	p-value
ช่วงเวลาเช้า	1	1,765	845	<0.001
	2	6		
ช่วงเวลากลางวัน	1	1,765	1	<0.001
	2	6		
ช่วงเวลาเย็น	1	1,765	1384.5	0.001
	2	6		
ช่วงเวลาดึก	1	1,765	4523.5	0.425^{ns}
	2	6		

หมายเหตุ : ns หมายถึง ไม่มีความแตกต่างกันทางสถิติ ที่ระดับนัยสำคัญ 0.05

เนื่องจากทำการทดสอบข้อกำหนดเบื้องต้นแล้ว พบว่า ข้อมูลไม่มีการแจกแจงปรกติ ดังตารางภาคผนวกที่ ค.1 ดังนั้น สถิติที่ใช้ในการทดสอบ คือ The Mann-Whitney U Test

จากตารางที่ 4.21 พบว่า ค่า U = 4523.5 และค่า p-value = 0.425 ซึ่งมีค่ามากกว่า 0.05 จึงยอมรับ H_0 ที่ระดับนัยสำคัญ 0.05 นั่นคือ จำนวนการเข้ามาใช้งานในช่วงเวลาดึกมีค่ามัธยฐานไม่แตกต่างกัน ระหว่างกลุ่มที่ 1 และกลุ่มที่ 2 ดังนั้นตัวแปรช่วงเวลาดึกจึงไม่สามารถนำมาจัดกลุ่มได้

4.4.1.2 การทดสอบสมมติฐานจากการจัดกลุ่มของวิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง

สมมติฐานการทดสอบ

$$H_0 : M_1 = M_2$$

$$H_1 : M_1 \neq M_2$$

ตารางที่ 4.22 ผลการทดสอบความแตกต่างระหว่างกลุ่มของวิธี OCSVM

ตัวแปร	กลุ่ม (Cluster)	n	Mann-Whitney U	p-value
ช่วงเวลาเช้า	1	2154	2462	0.008
	2	6		
ช่วงเวลากลางวัน	1	2154	1	<0.001
	2	6		

ตัวแปร	กลุ่ม (Cluster)	n	Mann-Whitney U	p-value
ช่วงเวลาเย็น	1	2154	1500	<0.001
	2	6		
ช่วงเวลาดึก	1	2154	5296	0.273^{ns}
	2	6		

หมายเหตุ : ns หมายถึง ไม่มีความแตกต่างกันทางสถิติ ที่ระดับนัยสำคัญ 0.05

เนื่องจากทำการทดสอบข้อกำหนดเบื้องต้นแล้ว พบว่า ข้อมูลไม่มีการแจกแจงปรกติ ดังตารางภาคผนวกที่ ค.2 ดังนั้น สถิติที่ใช้ในการทดสอบ คือ The Mann-Whitney U Test

จากตารางที่ 4.22 พบว่า ค่า U = 5296 และค่า p-value = 0.273 ซึ่งมีค่ามากกว่า 0.05 จึงยอมรับ H_0 ที่ระดับนัยสำคัญ 0.05 นั่นคือ จำนวนการเข้ามาใช้งานในช่วงเวลาดึก มีค่ามัธยฐานไม่แตกต่างกัน ระหว่างกลุ่มที่ 1 และกลุ่มที่ 2 ดังนั้นตัวแปรช่วงเวลาดึกจึงไม่สามารถนำมาจัดกลุ่มได้

4.4.1.3 การทดสอบสมมติฐานจากการจัดกลุ่มของวิธีไอโซเลชั่นฟอเรส

สมมติฐานการทดสอบ

$$H_0 : M_1 = M_2$$

$$H_1 : M_1 \neq M_2$$

ตารางที่ 4.23 ผลการทดสอบความแตกต่างระหว่างกลุ่มของวิธี Isolation Forest

ตัวแปร	กลุ่ม (Cluster)	n	Mann-Whitney U	p-value
ช่วงเวลาเช้า	1	2309	2276.5	0.004
	2	6		
ช่วงเวลากลางวัน	1	2309	1	<0.001
	2	6		
ช่วงเวลาเย็น	1	2309	1805	0.001
	2	6		
ช่วงเวลาดึก	1	2309	6280	0.636^{ns}
	2	6		

หมายเหตุ : ns หมายถึง ไม่มีความแตกต่างกันทางสถิติ ที่ระดับนัยสำคัญ 0.05

เนื่องจากทำการทดสอบข้อกำหนดเบื้องต้นแล้ว พบว่า ข้อมูลไม่มีการแจกแจงปรกติ ดังตารางภาคผนวกที่ ค.3 ดังนั้น สถิติที่ใช้ในการทดสอบ คือ The Mann-Whitney U Test

จากตารางที่ 4.23 พบว่า ค่า U = 6280 และค่า p-value = 0.636 ซึ่งมีค่ามากกว่า 0.05 จึงยอมรับ H_0 ที่ระดับนัยสำคัญ 0.05 นั่นคือ จำนวนการเข้ามาใช้งานในช่วงเวลาดึกมีค่ามัธยฐานไม่แตกต่างกัน ระหว่างกลุ่มที่ 1 และกลุ่มที่ 2 ดังนั้นตัวแปรช่วงเวลาดึกจึงไม่สามารถนำมาจัดกลุ่มได้

4.4.1.4 การทดสอบสมมติฐานจากการจัดกลุ่มของวิธีไอเอฟ-แอลโอเอฟ

สมมติฐานการทดสอบ

$$H_0 : M_1 = M_2$$

$$H_1 : M_1 \neq M_2$$

ตารางที่ 4.24 ผลการทดสอบความแตกต่างระหว่างกลุ่มของวิธี IF-LOF

ตัวแปร	กลุ่ม (Cluster)	n	Mann-Whitney U	p-value
ช่วงเวลาเช้า	1	222	208.5	0.003
	2	6		
ช่วงเวลากลางวัน	1	222	1	<0.001
	2	6		
ช่วงเวลาเย็น	1	222	172	0.001
	2	6		
ช่วงเวลาดึก	1	222	521.5	0.355^{ns}
	2	6		

หมายเหตุ : ns หมายถึง ไม่มีความแตกต่างกันทางสถิติ ที่ระดับนัยสำคัญ 0.05

เนื่องจากทำการทดสอบข้อกำหนดเบื้องต้นแล้ว พบว่า ข้อมูลไม่มีการแจกแจงปรกติ ดังตารางภาคผนวกที่ ค.4 ดังนั้น สถิติที่ใช้ในการทดสอบ คือ The Mann-Whitney U Test

จากตารางที่ 4.24 พบว่า ค่า U = 521.5 และค่า p-value = 0.3555 ซึ่งมีค่ามากกว่า 0.05 จึงยอมรับ H_0 ที่ระดับนัยสำคัญ 0.05 นั่นคือ จำนวนการเข้ามาใช้งานในช่วงเวลาดึกมีค่ามัธยฐานไม่แตกต่างกัน ระหว่างกลุ่มที่ 1 และกลุ่มที่ 2 ดังนั้นตัวแปรช่วงเวลาดึกจึงไม่สามารถนำมาจัดกลุ่มได้

4.4.2 การทดสอบสมมติฐานจากการจัดกลุ่มข้อมูลการเข้ามาใช้งานระบบที่มีสถานะล้มเหลว

4.4.2.1 การทดสอบสมมติฐานจากการจัดกลุ่มของวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์

สมมติฐานการทดสอบ

$$H_0 : M_1 = M_2$$

$$H_1 : M_1 \neq M_2$$

ตารางที่ 4.25 ผลการทดสอบความแตกต่างระหว่างกลุ่มของวิธี LOF

ตัวแปร	กลุ่ม (Cluster)	n	Mann-Whitney U	p-value
ช่วงเวลาเช้า	1	1064	6	<0.001
	2	6		
ช่วงเวลากลางวัน	1	1064	0	<0.001
	2	6		
ช่วงเวลาเย็น	1	1064	0	<0.001
	2	6		
ช่วงเวลาดึก	1	1064	5	<0.001
	2	6		

หมายเหตุ : ns หมายถึง ไม่มีความแตกต่างกันทางสถิติ ที่ระดับนัยสำคัญ 0.05

เนื่องจากทำการทดสอบข้อกำหนดเบื้องต้นแล้ว พบว่า ข้อมูลไม่มีการแจกแจงปรกติ ดังตารางภาคผนวกที่ ค.5 ดังนั้น สถิติที่ใช้ในการทดสอบ คือ The Mann-Whitney U Test

จากตารางที่ 4.25 พบว่าค่า p-value ในทุกช่วงเวลา <0.001 ซึ่งมีค่าน้อยกว่า 0.05 จึงปฏิเสธ H_0 ที่ระดับนัยสำคัญ 0.05 นั่นคือจำนวนการเข้ามาใช้งานในทุกช่วงเวลามีค่ามัธยฐานแตกต่างกัน ระหว่างกลุ่มที่ 1 และกลุ่มที่ 2

4.4.2.2 การทดสอบสมมติฐานจากการจัดกลุ่มของวิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง

สมมติฐานการทดสอบ

$$H_0 : M_1 = M_2$$

$$H_1 : M_1 \neq M_2$$

ตารางที่ 4.26 ผลการทดสอบความแตกต่างระหว่างกลุ่มของวิธี OCSVM

ตัวแปร	กลุ่ม (Cluster)	n	Mann-Whitney U	p-value
ช่วงเวลาเช้า	1	974	179	<0.001
	2	29		
ช่วงเวลากลางวัน	1	974	1679.5	<0.001
	2	29		
ช่วงเวลาเย็น	1	974	4314	<0.001
	2	29		
ช่วงเวลาดึก	1	974	2753.5	<0.001
	2	29		

หมายเหตุ : ns หมายถึง ไม่มีความแตกต่างกันทางสถิติ ที่ระดับนัยสำคัญ 0.05

เนื่องจากทำการทดสอบข้อกำหนดเบื้องต้นแล้ว พบว่า ข้อมูลไม่มีการแจกแจงปรกติ ดังตารางภาคผนวกที่ ค.6 ดังนั้น สถิติที่ใช้ในการทดสอบ คือ The Mann-Whitney U Test

จากตารางที่ 4.26 พบว่าค่า p-value ในทุกช่วงเวลา <0.001 ซึ่งมีค่าน้อยกว่า 0.05 จึงปฏิเสธ H_0 ที่ระดับนัยสำคัญ 0.05 นั่นคือจำนวนการเข้ามาใช้งานในทุกช่วงเวลามีค่ามัธยฐานแตกต่างกัน ระหว่างกลุ่มที่ 1 และกลุ่มที่ 2

4.4.2.3 การทดสอบสมมติฐานจากการจัดกลุ่มของวิธีไอโซเลชั่นฟอเรส

สมมติฐานการทดสอบ

$$H_0 : M_1 = M_2$$

$$H_1 : M_1 \neq M_2$$

ตารางที่ 4.27 ผลการทดสอบความแตกต่างระหว่างกลุ่มของวิธี Isolation Forest

ตัวแปร	กลุ่ม (Cluster)	n	Mann-Whitney U	p-value
ช่วงเวลาเช้า	1	1052	179	<0.001
	2	29		
ช่วงเวลากลางวัน	1	1052	1986.5	<0.001
	2	29		
ช่วงเวลาเย็น	1	1052	5397	<0.001
	2	29		
ช่วงเวลาดึก	1	1052	3529.5	<0.001
	2	29		

หมายเหตุ : ns หมายถึง ไม่มีความแตกต่างกันทางสถิติ ที่ระดับนัยสำคัญ 0.05

เนื่องจากทำการทดสอบข้อกำหนดเบื้องต้นแล้ว พบว่า ข้อมูลไม่มีการแจกแจงปรกติ ดังตารางภาคผนวกที่ ค.7 ดังนั้น สถิติที่ใช้ในการทดสอบ คือ The Mann-Whitney U Test

จากตารางที่ 4.27 พบว่าค่า p-value ในทุกช่วงเวลา <0.001 ซึ่งมีค่าน้อยกว่า 0.05 จึงปฏิเสธ H_0 ที่ระดับนัยสำคัญ 0.05 นั่นคือจำนวนการเข้ามาใช้งานในทุกช่วงเวลามีค่ามัธยฐานแตกต่างกัน ระหว่างกลุ่มที่ 1 และกลุ่มที่ 2

4.4.2.4 การทดสอบสมมติฐานจากการจัดกลุ่มของวิธีไอเอฟ-แอลโอเอฟ

สมมติฐานการทดสอบ

$$H_0 : M_1 = M_2$$

$$H_1 : M_1 \neq M_2$$

ตารางที่ 4.28 ผลการทดสอบความแตกต่างระหว่างกลุ่มของวิธี IF-LOF

ตัวแปร	กลุ่ม (Cluster)	n	Mann-Whitney U	p-value
ช่วงเวลาเช้า	1	161	6	<0.001
	2	6		
ช่วงเวลากลางวัน	1	161	0	<0.001
	2	6		
ช่วงเวลาเย็น	1	161	0	<0.001
	2	6		
ช่วงเวลาดึก	1	161	5	<0.001
	2	6		

หมายเหตุ : ns หมายถึง ไม่มีความแตกต่างกันทางสถิติ ที่ระดับนัยสำคัญ 0.05

เนื่องจากทำการทดสอบข้อกำหนดเบื้องต้นแล้ว พบว่า ข้อมูลไม่มีการแจกแจงปรกติ ดังตารางภาคผนวกที่ ค.8 ดังนั้น สถิติที่ใช้ในการทดสอบ คือ The Mann-Whitney U Test

จากตารางที่ 4.28 พบว่าค่า p-value ในทุกช่วงเวลา <0.001 ซึ่งมีค่าน้อยกว่า 0.05 จึงปฏิเสธ H_0 ที่ระดับนัยสำคัญ 0.05 นั่นคือจำนวนการเข้ามาใช้งานในทุกช่วงเวลามีค่ามัธยฐานแตกต่างกัน ระหว่างกลุ่มที่ 1 และกลุ่มที่ 2

4.5. อภิปรายผลการวิจัย

จากการตรวจจับสิ่งผิดปรกติด้วยวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง วิธีไอโซเลชั่นฟอเรส และวิธีไอเอฟ-แอลโอเอฟ ร่วมกับข้อมูล 2 ชุด ซึ่งเป็นข้อมูลที่เก็บเฉพาะสถานะการเข้ามาใช้งานในระบบสำเร็จ และสถานะการเข้ามาใช้งานในระบบล้มเหลว พบว่า วิธีการตรวจจับสิ่งผิดปรกติด้วยวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง และวิธีไอโซเลชั่นฟอเรส มีสัดส่วนในการตรวจจับสิ่งผิดปรกติที่ใกล้เคียงกัน โดยงานวิจัยของ Henriksson (2021) กล่าวว่า วิธีการตรวจจับสิ่งผิดปรกติด้วยวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง และวิธีไอโซเลชั่นฟอเรส มีสัดส่วนค่าความถ่วงดุล ค่าความแม่น และค่าความระลึกใกล้เคียงกันทั้ง 3 วิธี ในส่วนของวิธีไอเอฟ-แอลโอเอฟ สามารถตรวจจับสิ่งผิดปรกติได้สัดส่วนน้อยที่สุดเมื่อเทียบกับอีก 3 วิธี ซึ่งสามารถตรวจจับรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จ คิดเป็นร้อยละ 0.19 และสามารถตรวจจับรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลว คิดเป็นร้อยละ 0.30 ซึ่งแตกต่างจากงานวิจัยของ Zhangyu et al. (2019) ที่กล่าวว่า การตรวจจับสิ่งผิดปรกติของวิธีไอเอฟ-แอลโอเอฟมีสัดส่วนค่าความถ่วงดุล และค่าความแม่นสูงที่สุด สาเหตุมาจากวิธีไอเอฟ-แอลโอเอฟ เป็นแบบจำลองที่มีลักษณะผสมผสานกันระหว่างไอโซเลชั่นฟอเรส และวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ ซึ่งในการตรวจจับนั้นเป็นการตรวจจับแบบ 2 ขั้นตอน คือ ขั้นตอนที่ 1 ทำการตรวจจับสิ่งผิดปรกติโดยวิธีไอโซเลชั่นฟอเรส จากนั้นขั้นตอนที่ 2 เป็นการนำสิ่งผิดปรกติที่ได้จากการตรวจจับด้วยวิธีไอโซเลชั่นฟอเรส มาทำการตรวจจับด้วยวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ ทำให้ได้ผลลัพธ์จากการตรวจจับสิ่งผิดปรกติของวิธีไอเอฟ-แอลโอเอฟออกมาในสัดส่วนที่ต่ำ

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

สรุปผลการวิจัยการตรวจจับสิ่งผิดปรกติของการเข้าใช้ระบบ Azure Active Directory ด้วยเทคนิคการเรียนรู้ของเครื่อง ซึ่งมีวัตถุประสงค์ดังนี้ 1) เพื่อศึกษาวิธีการตรวจจับสิ่งผิดปรกติของการเข้ามาใช้งานในระบบด้วยเทคนิคการตรวจจับสิ่งผิดปรกติ 2) วิเคราะห์ผลการตรวจจับสิ่งผิดปรกติของการเข้าใช้งานในระบบโดยอาศัยวิธีการจัดกลุ่ม และสถิติเชิงอนุมาน สรุปผลได้ดังนี้

5.1 สรุปผลการวิจัย

5.1.1 ข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จ

ผู้วิจัยทำการตรวจจับสิ่งผิดปรกติของการเข้าใช้ระบบมีสถานะสำเร็จ สำหรับเป็นแนวทางเบื้องต้นในการตรวจสอบสิ่งผิดปรกติของการเข้าใช้ระบบมีสถานะสำเร็จ

วิธีการตรวจจับสิ่งผิดปรกติ จำนวนทั้งหมด 4 วิธี ได้แก่ วิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ตรวจจับสิ่งผิดปรกติได้ 1,771 รายการ วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่งตรวจจับสิ่งผิดปรกติได้ 2,160 รายการ วิธีไอโซเลชั่นฟอเรสตรวจจับสิ่งผิดปรกติได้ 2,160 รายการ และวิธีไอเอฟ-แอลโอเอฟตรวจจับสิ่งผิดปรกติได้ 228 รายการ จากนั้นนำสิ่งผิดปรกติที่ตรวจจับได้ในแต่ละวิธีไปทำการจัดกลุ่มด้วยเทคนิคเคมีน ผลพบว่าทั้ง 4 วิธี ให้ผลลัพธ์ในการจัดกลุ่มออกมาเหมือนกัน คือ จำนวน 2 กลุ่ม โดยกลุ่มที่ 1 จะมีจำนวนของสิ่งผิดปรกติเป็นจำนวนมาก ซึ่งจะพฤติกรรมการเข้ามาใช้งานโดยเฉลี่ยไม่แตกต่างกันมากนักเนื่องจากมีการกระจายตัวค่อนข้างต่ำ ในส่วนของกลุ่มที่ 2 จะพบว่าเป็นกลุ่มของสิ่งผิดปรกติที่มีพฤติกรรมการเข้ามาใช้งานเฉลี่ยเป็นอย่างมากอย่างเห็นได้ชัดเจน และมีการกระจายตัวจากข้อมูลในกลุ่มแรกอย่างชัดเจนซึ่งเป็นกลุ่มที่ควรได้รับการตรวจสอบเป็นอันดับแรก และได้นำผลการจัดกลุ่มไปทดสอบเพื่อตรวจว่าคุณลักษณะ หรือตัวแปรสามารถแบ่งค่าความแตกต่างระหว่างกลุ่มของการเข้ามาใช้งานระบบในแต่ละช่วงเวลา โดยใช้สถิติทดสอบ Mann-Whitney U และผลที่ได้จากการวิเคราะห์คือค่าเฉลี่ยของตัวแปรช่วงเวลาเช้า ช่วงเวลากลางวัน และช่วงเวลาเย็นสำหรับกลุ่มที่ 1 แตกต่างจากกลุ่มที่ 2 ในขณะที่ค่าเฉลี่ยของตัวแปรช่วงเวลาดึกสำหรับกลุ่มที่ 1 ไม่แตกต่างจากกลุ่มที่ 2 นั่นคือ 3 ตัวแปรได้แก่ ตัวแปรช่วงเวลาเช้า ช่วงเวลากลางวัน และช่วงเวลาเย็น มีความเหมาะสมที่จะนำไปใช้ในการจัดกลุ่ม

5.1.2 ข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลว

ผู้วิจัยทำการตรวจจับสิ่งผิดปรกติของการเข้าใช้ระบบมีสถานะล้มเหลว สำหรับเป็นแนวทางเบื้องต้นในการตรวจสอบสิ่งผิดปรกติของการเข้าใช้ระบบมีสถานะล้มเหลว

วิธีการตรวจจับสิ่งผิดปรกติจำนวนทั้งหมด 4 วิธี ได้แก่ วิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ตรวจจับสิ่งผิดปรกติได้ 1,070 รายการ วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่งตรวจจับสิ่งผิดปรกติได้ 1,003 รายการ วิธีไอโซเลชั่นฟอเรสตรวจจับสิ่งผิดปรกติได้ 1,081 รายการ และวิธีไอเอฟ-แอลโอเอฟตรวจจับสิ่งผิดปรกติได้ 167 รายการ จากนั้นนำสิ่งผิดปรกติที่ตรวจจับได้ในแต่ละวิธีไปทำการจัดกลุ่มด้วยเทคนิคเคมีน ผลพบว่าทั้ง 4 วิธี ให้ผลลัพธ์ในการจัดกลุ่มออกมาเหมือนกัน คือ จำนวน 2 กลุ่ม โดยกลุ่มที่ 1 จะมีจำนวนของสิ่งผิดปรกติเป็นจำนวนมาก ซึ่งจะพฤติกรรมการเข้า

มาใช้งานโดยเฉลี่ยไม่แตกต่างกันมากนักเนื่องจากมีการกระจายตัวค่อนข้างต่ำ ในส่วนของกลุ่มที่ 2 จะพบว่าเป็นกลุ่มของสิ่งผิดปรกติที่มีพฤติกรรมการเข้ามาใช้งานเฉลี่ยเป็นอย่างมากอย่างเห็นได้ชัดเจน และมีการกระจายตัวจากข้อมูลในกลุ่มแรกอย่างชัดเจนซึ่งเป็นกลุ่มที่ควรได้รับการตรวจสอบเป็นอันดับแรก และได้นำผลการจัดกลุ่มไปทดสอบเพื่อตรวจว่าคุณลักษณะหรือตัวแปรสามารถแบ่งค่าความแตกต่างระหว่างกลุ่มของการเข้ามาใช้งานระบบในแต่ละช่วงเวลา โดยใช้สถิติทดสอบ Mann-Whitney U และผลที่ได้จากการวิเคราะห์คือทุกตัวแปรสามารถใช้ในการจัดกลุ่มของข้อมูลความถี่การเข้าใช้ระบบที่มีสถานะล้มเหลว

5.2 ข้อจำกัดและข้อเสนอแนะ

5.2.1 ข้อจำกัด

1) เนื่องจากเวลาในการทำวิจัยที่จำกัด จึงไม่สามารถนำวิธีการตรวจจับสิ่งผิดปรกติวิธีอื่นมาทำการทดสอบได้ซึ่งอาจจะมีวิธีที่ให้ประสิทธิภาพการทำนายที่ดีกว่า
2) เนื่องด้วยข้อจำกัดเกี่ยวกับอุปกรณ์และเครื่องมือที่ใช้ในการวิเคราะห์ ทำให้ไม่สามารถปรับแต่งค่าพารามิเตอร์ที่เหมาะสมสำหรับวิธีการตรวจจับสิ่งผิดปรกติแต่ละวิธี

5.2.2 ข้อเสนอแนะ

1) พิจารณาวิธีการตรวจจับสิ่งผิดปรกติแบบไม่มีผู้สอนอื่นๆ ได้แก่ Histogram-based Outlier Score (HBOS) เพื่อไปใช้ประโยชน์เมื่อต้องการทราบความผิดปรกติของข้อมูลที่ไม่มีผลเฉลย
2) ควรมีการพิจารณาตัวแปรอื่น ๆ เพิ่มเติมสำหรับตรวจจับสิ่งผิดปรกติ เช่น ประเภทแอปพลิเคชั่น ประเทศ และอุปกรณ์ที่ใช้งาน

5.3 แนวทางที่จะศึกษาต่อในอนาคต

ทางผู้วิจัยทำการนำผลลัพธ์ของการตรวจจับสิ่งผิดปรกติจากการศึกษาวิจัยในครั้งนี้ ส่งต่อให้ทางฝ่าย Cyber Security ของบริษัท จากนั้นทางบริษัทจะมีการประเมินสิ่งผิดปรกติและกำหนดป้ายกำกับคำตอบของข้อมูลโดยผู้เชี่ยวชาญ ซึ่งการกำหนดป้ายกำกับคำตอบนี้ ส่งผลให้ในอนาคตสามารถใช้วิธีการตรวจจับสิ่งผิดปรกติด้วยวิธีการเรียนรู้แบบมีผู้สอนได้ เช่น K -Nearest Neighbor (k-NN), Bayesian Network (BN), Supervised Neural Network (NN), Decision Tree (DT) ดังนั้นผู้วิจัยจึงมีความสนใจที่จะทำการตรวจจับสิ่งผิดปรกติของการเข้าใช้ระบบต่อไป โดยใช้วิธีการเรียนรู้แบบมีผู้สอน ซึ่งวิธีการเรียนรู้แบบมีผู้สอนสามารถวัดประสิทธิภาพได้อย่างเป็นรูปธรรมมากขึ้น เช่น การวัดประสิทธิภาพด้วยค่าความแม่น ค่าความระลึก หรือค่าความเที่ยง เป็นต้น

เอกสารอ้างอิง

กัลยา วานิชย์บัญชา. 2552. **การวิเคราะห์ข้อมูลหลายตัวแปร.** พิมพ์ครั้งที่ 4. กรุงเทพฯ : ธรรมสาร

ชาคริต. 2563 **Anomaly คืออะไร**. [ออนไลน์]. เข้าถึงได้จาก https://www.softnix.co.th/2020/02/ anomaly-detection-part-1

ณิชาภา อยู่ผ่องภา. 2565. **สถิติภัยคุกคามครึ่งปีแรก 2022 จากศูนย์ CSOC ของ NT cyfence**. [ออนไลน์]. เข้าถึงได้จาก https://www.cyfence.com/article/first-half-2022-threat-statistics-from-nt-csoc/

ธนชัย จิระจันทร์. 2557. **การปรับปรุงประสิทธิภาพของการตรวจจับสิ่งผิดปรกติสำหรับการวิเคราะห์ปูมแบบปรับขนาดได้.** วิศวกรรมศาสตรมหาบัณฑิต. สาขาวิชาวิศวกรรมคอมพิวเตอร์. จุฬาลงกรณ์มหาวิทยาลัย

ปาริชาติ โพธิอินทร์. 2564. **Azure Active Directory คืออะไร ?**. [ออนไลน์]. เข้าถึงได้จาก https://monsterconnect.co.th/azure-ad-vs-ad-ds/

พิพัฒน์ สมโลก. 2563. **Machine Learning สิ่งใกล้ตัวแห่งโลกยุคใหม่.** [ออนไลน์]. เข้าถึงได้จาก https://www.depa.or.th/th/article-view/article11-2563

วีระศักดิ์ จินารัตน์. 2557. **การวิเคราะห์ข้อมูลด้วยสถิติเชิงอนุมาน**. วารสารวิชาการมหาวิทยาลัยการจัดการและเทคโนโลยี. 11(2): 80-84.

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

ศศิวุฒิ ชัยเดชา. 2563. **หาจำนวน Clusters ที่เหมาะสมสำหรับ KMeans clustering ด้วย Elbow method**. [ออนไลน์]. เข้าถึงได้จาก : https://lengyi.medium.com/หาจำนวน-clusters-ที่เหมาะสม-kmeans-clustering-ด้วย-elbow-method-85421efe9d

ศุภโชคชัย แซ่ตัน. 2564. **การประเมินระดับความเสี่ยงด้านไซเบอร์และการบริหารความเสี่ยงด้านไซเบอร์ภายธุรกิจและองค์กร**. [ออนไลน์]. เข้าถึงได้จาก :https://www.depa.or.th/th/article-view/cyber-risk-assessment-and-cyber-risk-management

ศุภามณ จันทร์สกุล และสุชาดา บวรกิติวงศ์. 2560. **สถิตินอนพาราเมตริกและการประยุกต์ใช้ในงานวิจัยทางการแพทย์**. EAU HERITAGE JOURNAL Science and Technology. 11(1) : 39-42.

เอกสารอ้างอิง(ต่อ)

สมชาย อารยพิทยา. 2554. **ข้อมูลจราจรคอมพิวเตอร์(logfile)** . [ออนไลน์]. เข้าถึงได้จาก https://erp.mju.ac.th/blog.aspx?bid=437

สุรวัชร ศรีเปารยะ และสายชล สินสมบูรณ์ทอง. 2560. **การเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มการเป็นโรคไตเรื้อรัง : กรณีศึกษาโรงพยาบาลแห่งหนึ่งในประเทศอินเดีย**. วารสารวิทยาศาสตร์และเทคโนโลยี. 25(5). 839-853

สายชล สินสมบูรณ์ทอง. 2563. **สถิติไม่อิงพารามิเตอร์**. พิมพ์ครั้งที่ 2. กรุงเทพฯ : จามจุรีโปรดักส์.

สำนักงานรัฐบาลอิเล็กทรอนิกส์. 2559**. ความมั่นคงปลอดภัยทางไซเบอร์ (Cyber Security) คืออะไร**. [ออนไลน์]. เข้าถึงได้จาก http://www.takesa2.go.th//download/learn_online/cyber_doc.pdf

สุทิน ชนะบุญ. 2560. **สถิติและการวิเคราะห์ข้อมูลในงานวิจัยเบื้องต้น**. ขอนแก่น : สำนักงานสาธารณสุขจังหวัดขอนแก่น

สุจิตรา สุคนธมัต 2565. **เอกสารประกอบการสอบวิชาโปรแกรมสำเร็จรูปทางสถิติ**. ภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

สุพัฒน์ สุกมลสันต์. 2560. **การเปรียบเทียบก่อนและหลังการทดสอบรวมเพื่อการวิจัย**. วารสารวิชาการ มหาวิทยาลัยราชภัฏบุรีรัมย์. 9(2): 51-69.

สุดา เธียรมนตรี. 2563. **คู่มือเรียนเขียนโปรแกรมภาษา Python ฉบับสมบูรณ์**. พิมพ์ครั้งที่ 1. นนทบุรี: ไอดีซี

อุมาพร จันทศร 2542. **สถิติที่ไม่ใช้พารามิเตอร์** สำนักพิมพ์ฟิสิกส์เซ็นเตอร์ กรุงเทพฯ

อาณัติชัย เตชะวิเศษชัย. 2565 . **Anomaly detection with Isolation forest: แยกข้อมูลผิดปรกติง่ายๆ ด้วย Isolation Forest**. [ออนไลน์]. เข้าถึงได้จาก : https://bigdataexperience.org/anomaly-detection-with-isolation-forest.

อรพิน ประวัติบริสุทธิ์. 2564. **Python สำหรับงาน Data Science Data Visualization และ Machine Learning**. พิมพ์ครั้งที่ 1. กรุงเทพฯ : โปรวิชั่น

Agrawal, S. and Jitendra, A. 2015. **Survey on anomaly detection using data mining techniques**. Procedia Computer Science. 60: p. 708-713.

Amy. 2021. **One-class support vector machine (SVM) for anomaly detection**. [Online]. Available https://grabngoinfo.com/one-class-support-vector-machine-svm-for-anomaly-detection/

เอกสารอ้างอิง(ต่อ)

Amer, M. 2013. **Enhancing one-class support vector machines for unsupervised anomaly detection**. Faculty of Postgraduate Studies and Scientific Research German University in Cairo.

Breunig, M., Kriegel, H., Ng, R and Sander, J. 2000 **LOF: Identifying density-based local outliers**. Proc. ACM SIGMOD 2000 Int. Conf. On Management of Data, Dalles, TX, 2000.

Baddar, S., Merlo, A., Migliardi, M. 2014**. Anomaly Detection in Computer Networks: A State of the Art Review**. Journal of Wireless Mobile Networks. Ubiquitous Computing and Dependable Applications. pp. 29-64

Campos, G.O. et al. 2016. **On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study**. Data Min Knowl Disc 30 : 891–927 https://doi.org/10.1007/s10618-015-0444-8

Chandola, V., Banerjee, A. and Kumar, V. (2009) **Anomaly detection: A survey**. ACM Computing Surveys 41, pp. 1–58. URL: https://doi.org/10.1145/1541880.1541882.

Deza, E. and Deza, M. 2009**. Encyclopedia of Distances**. (94). Heidelberg : Springer Berlin.: 323-324.

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.

Forsmark, M. 2020. **Anomaly Detection in User Authentication Logs using Long Short-Term Memories and Word Embeddings**. Degree Project In Computer science and engineering Second Cycle Kth Royal Institute Of Technology School Of Electrical Engineering And Computer Sciences

Goldstein, M. and Uchida, S. 2016. **A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data**. PLOS ONE. 11(4):e0152173. https://doi.org/10.1371/journal.pone.0152173.

Hawkins D.1980. **Identification of outliers**, vol. 13. Netherlands. Springer.

Hartigan, J. and Wong, M. 1979. **A K-means clustering algorithm**. Journal of the Royal Statistical Society Series C (Applied Statistics). 28: 100-108.

เอกสารอ้างอิง(ต่อ)

Hsu, C., Chang, C. and Lin, C. 2016. **A practical guide to support vector classification**. Department of Computer Science National Taiwan University.

Hui, C. 2021. **Anomaly Detection Analysis - Isolation Forest**. [Online]. Available. https://deepnote.com/@christopher-hui/Anomaly-Detection-Analysis-Isolation-Forest-c012da68-8081-4e2e-9bc8-8bc59a1c2d6c

Henriksson, A. 2021. **Unsupervised Anomaly Detection in Time Series Data An Implementation on Electricity Consumption Series**. Degree Project In Mathematics Second Cycle Kth Royal Institute Of Technology School Of Engineering Science

Jääskelä, J. 2020. **Anomaly-Based Insider Threat Detection with Expert Feedback and Descriptions.** University of Oulu Degree Programme in Computer Science and Engineering

Kumar, V. 2002. **Parallel Issues in Data Mining**. VECPAR.

Krause, E. 1987. **Taxicab Geometry: An adventure in non-Euclidean geometry**. USA : Dover

Myatt, G. and Johnson, W. 2009. **Making Sense of Data II: A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications**. Journal of Statistical Software. 34(1).

Patcha, A. and Park, J. 2007. **An overview of anomaly detection techniques: existing solutions and latest technological trends. Computer Networks**. 51: 3448–3470.

Rajasegarar, S., Leckie, C., and Palaniswami, M. (2008). **Anomaly detection in wireless sensor networks**. In Proceedings of the IEEE Wireless Communications. 15 (4): pp. 34-40.

Rousseeuw, P. 1987. **Silhouettes: a graphical aid to the interpretation and validation of cluster analysis**. Journal of Computational and Applied Mathematics. 53-65.

Tran, Q., Duan, H. and Li, X. 2004. **One-class Support Vector Machine for Anomaly Network Traffic Detection**. China Education and Research Network (CERNET) Tsinghua University, China

เอกสารอ้างอิง(ต่อ)

Wilcox, R. R. (1990). **Comparing the mean of two independent group**. Biometrical Journal. 32. 771-780.

Wilcox, R.R. 2012. **Modern Statistics for the Social and Behavioral Sciences: A Practical Introduction**. Boca Raton: Chapman & Hall.

Zhang, R. et al. 2007. **One Class Support Vector Machine for Anomaly Detection in the Communication Network Performance Data**. 5th WSEAS Int. Conference on Applied Electromagnetics, Wireless and Optical Communications, Spain.

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 ชุดคำสั่งไพทอน (Python) ที่ใช้ในการเก็บรวบรวมข้อมูล และการสร้างวิธีการตรวจจับสิ่งผิดปรกติ เพื่อทำให้เกิดความเข้าใจถึงพฤติกรรมและลักษณะของเหตุการณ์ที่ผิดปรกติของการเข้ามาใช้งานระบบ

ภาคผนวก ก.1 ชุดคำสั่งไพทอนในการเตรียมข้อมูลรายการเข้าใช้ระบบที่มีสถานะสำเร็จ

```
# Importing the libraries

import pandas as pd
from datetime import datetime, date
from numpy import nan
import numpy as np
AzureActiveDirectory_Logfail=pd.read_csv('ตำแหน่งที่เก็บไฟล์')
AzureActiveDirectory_Logfail
# Check Null
AzureActiveDirectory_Logfail.isnull().sum()
# Check null in column email
AzureActiveDirectory_Logfail[AzureActiveDirectory_Logfail['user_email'].isnull()]
#Build DataFrame
AzureActiveDirectory_Logfail['count_login'] = 0
AzureActiveDirectory_Logfail_1H
```

```
#Convert timestamp to data hour time
AzureActiveDirectory_Logfail['date_timestamp'] =
pd.to_datetime(AzureActiveDirectory_Logfail['date_timestamp'])
AzureActiveDirectory_Logfail['date'] = AzureActiveDirectory_Logfail['date_timestamp'].dt.date
AzureActiveDirectory_Logfail['weekday'] = AzureActiveDirectory_Logfail['timestamp'].apply(lambda
x: x.weekday())
AzureActiveDirectory_Logfail['hour'] =
AzureActiveDirectory_Logfail['date_timestamp'].apply(lambda x: x.hour)
AzureActiveDirectory_Logfail['time'] =
AzureActiveDirectory_Logfail['date_timestamp'].dt.strftime('%H:%M:%S')
AzureActiveDirectory_Logfail
#Groupby column and count value login
AzureActiveDirectory_Logfail_1H =
AzureActiveDirectory_Logfail.groupby(['date','user_email','audit_Operation','hour'])['count_login'].co
unt().reset_index()
AzureActiveDirectory_Logfail_1H
```

```
# pivot the table
pivot_AzureActiveDirectory_Logfail_1H =
AzureActiveDirectory_Logfail_1H.pivot_table(values='count_login', index=['date','user_email'],
columns='hour', aggfunc='sum')
pivot_AzureActiveDirectory_Logfail_1H
#Rename Column
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={0: "00.00AM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={1: "01.00AM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={2: "02.00AM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={3: "03.00AM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={4: "04.00AM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={5: "05.00AM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={6: "06.00AM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={7: "07.00AM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={8: "08.00AM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={9: "09.00AM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={10: "10.00AM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={11: "11.00AM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={12: "12.00PM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={13: "13.00PM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={14: "14.00PM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={15: "15.00PM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={16: "16.00PM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={17: "17.00PM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={18: "18.00PM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={19: "19.00PM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={20: "20.00PM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={21: "21.00PM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={22: "22.00PM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={23: "23.00PM"},inplace=True)
```

```
#fill NaN
pivot_AzureActiveDirectory_Logfail_1H = pivot_AzureActiveDirectory_Logfail_1H.fillna(value=0)
pivot_AzureActiveDirectory_Logfail_1H
#Build Column To Defind time period
pivot_AzureActiveDirectory_Logfail_1H['Morning'] = 0 #06.00-11.00
pivot_AzureActiveDirectory_Logfail_1H['Afternoon'] = 0 #12.00-17.00
pivot_AzureActiveDirectory_Logfail_1H['Evening'] = 0 #18.00-23.00
pivot_AzureActiveDirectory_Logfail_1H['Night'] = 0 #00.00-05.00
#Sum value
pivot_AzureActiveDirectory_Logfail_1H['Morning']=pivot_AzureActiveDirectory_Logfail_1H['06.00AM'
]+pivot_AzureActiveDirectory_Logfail_1H['07.00AM']+pivot_AzureActiveDirectory_Logfail_1H['08.00A
M']+pivot_AzureActiveDirectory_Logfail_1H['09.00AM']+pivot_AzureActiveDirectory_Logfail_1H['10.0
0AM']+pivot_AzureActiveDirectory_Logfail_1H['11.00AM']
pivot_AzureActiveDirectory_Logfail_1H['Afternoon']=pivot_AzureActiveDirectory_Logfail_1H['12.00P
M']
+pivot_AzureActiveDirectory_Logfail_1H['13.00PM']+pivot_AzureActiveDirectory_Logfail_1H['14.00P
M']+pivot_AzureActiveDirectory_Logfail_1H['15.00PM']+pivot_AzureActiveDirectory_Logfail_1H['16.0
0PM']+pivot_AzureActiveDirectory_Logfail_1H['17.00PM']
pivot_AzureActiveDirectory_Logfail_1H['Evening']=pivot_AzureActiveDirectory_Logfail_1H['18.00PM']
+pivot_AzureActiveDirectory_Logfail_1H['19.00PM']+pivot_AzureActiveDirectory_Logfail_1H['20.00P
M']+pivot_AzureActiveDirectory_Logfail_1H['21.00PM']+pivot_AzureActiveDirectory_Logfail_1H['22.0
0PM']+pivot_AzureActiveDirectory_Logfail_1H['23.00PM']
pivot_AzureActiveDirectory_Logfail_1H['Night']=pivot_AzureActiveDirectory_Logfail_1H['00.00AM']+
pivot_AzureActiveDirectory_Logfail_1H['01.00AM']+pivot_AzureActiveDirectory_Logfail_1H['02.00AM
']+pivot_AzureActiveDirectory_Logfail_1H['03.00AM']+pivot_AzureActiveDirectory_Logfail_1H['04.00
AM']+pivot_AzureActiveDirectory_Logfail_1H['05.00AM']
#reset index
rein_pivot_AzureActiveDirectory_Logfail_1H=pivot_AzureActiveDirectory_Logfail_1H.reset_index()
rein_pivot_AzureActiveDirectory_Logfail_1H
```

ภาคผนวก ก.2 ชุดคำสั่งไพทอนในการเตรียมข้อมูลรายการเข้าใช้ระบบที่มีสถานะล้มเหลว

```
# Importing the libraries

import pandas as pd
from datetime import datetime, date
from numpy import nan
import numpy as np
AzureActiveDirectory_Logfail=pd.read_csv('ตำแหน่งที่เก็บไฟล์')
# Check Null
AzureActiveDirectory_Logfail.isnull().sum()
# Check null in column email
AzureActiveDirectory_Logfail[AzureActiveDirectory_Logfail['user_email'].isnull()]
#Build DataFEame
AzureActiveDirectory_Logfail['count_login'] = 0
```

```
#Convert timestamp to data hour time
AzureActiveDirectory_Logfail['date_timestamp'] =
pd.to_datetime(AzureActiveDirectory_Logfail['date_timestamp'])
AzureActiveDirectory_Logfail['date'] = AzureActiveDirectory_Logfail['date_timestamp'].dt.date
AzureActiveDirectory_Logfail['weekday'] = AzureActiveDirectory_Logfail['timestamp'].apply(lambda
x: x.weekday())
AzureActiveDirectory_Logfail['hour'] =
AzureActiveDirectory_Logfail['date_timestamp'].apply(lambda x: x.hour)
AzureActiveDirectory_Logfail['time'] =
AzureActiveDirectory_Logfail['date_timestamp'].dt.strftime('%H:%M:%S')
AzureActiveDirectory_Logfail
#Groupby column and count value login
AzureActiveDirectory_Logfail_1H =
AzureActiveDirectory_Logfail.groupby(['date','user_email','audit_Operation','hour'])['count_login'].co
unt().reset_index()
AzureActiveDirectory_Logfail_1H
```

```
# pivot the table
pivot_AzureActiveDirectory_Logfail_1H =
AzureActiveDirectory_Logfail_1H.pivot_table(values='count_login', index=['date','user_email'],
columns='hour', aggfunc='sum')
pivot_AzureActiveDirectory_Logfail_1H
#Rename Column
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={0: "00.00AM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={1: "01.00AM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={2: "02.00AM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={3: "03.00AM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={4: "04.00AM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={5: "05.00AM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={6: "06.00AM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={7: "07.00AM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={8: "08.00AM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={9: "09.00AM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={10: "10.00AM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={11: "11.00AM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={12: "12.00PM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={13: "13.00PM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={14: "14.00PM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={15: "15.00PM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={16: "16.00PM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={17: "17.00PM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={18: "18.00PM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={19: "19.00PM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={20: "20.00PM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={21: "21.00PM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={22: "22.00PM"},inplace=True)
pivot_AzureActiveDirectory_Logfail_1H.rename(columns={23: "23.00PM"},inplace=True)
```

```
#fill NaN
pivot_AzureActiveDirectory_Logfail_1H = pivot_AzureActiveDirectory_Logfail_1H.fillna(value=0)
pivot_AzureActiveDirectory_Logfail_1H
#Build Column To Defind time period
pivot_AzureActiveDirectory_Logfail_1H['Morning'] = 0 #06.00-11.00
pivot_AzureActiveDirectory_Logfail_1H['Afternoon'] = 0 #12.00-17.00
pivot_AzureActiveDirectory_Logfail_1H['Evening'] = 0 #18.00-23.00
pivot_AzureActiveDirectory_Logfail_1H['Night'] = 0 #00.00-05.00
#Sum value
pivot_AzureActiveDirectory_Logfail_1H['Morning']=pivot_AzureActiveDirectory_Logfail_1H['06.00AM'
]+pivot_AzureActiveDirectory_Logfail_1H['07.00AM']+pivot_AzureActiveDirectory_Logfail_1H['08.00A
M']+pivot_AzureActiveDirectory_Logfail_1H['09.00AM']+pivot_AzureActiveDirectory_Logfail_1H['10.0
0AM']+pivot_AzureActiveDirectory_Logfail_1H['11.00AM']
pivot_AzureActiveDirectory_Logfail_1H['Afternoon']=pivot_AzureActiveDirectory_Logfail_1H['12.00P
M']
+pivot_AzureActiveDirectory_Logfail_1H['13.00PM']+pivot_AzureActiveDirectory_Logfail_1H['14.00P
M']+pivot_AzureActiveDirectory_Logfail_1H['15.00PM']+pivot_AzureActiveDirectory_Logfail_1H['16.0
0PM']+pivot_AzureActiveDirectory_Logfail_1H['17.00PM']
pivot_AzureActiveDirectory_Logfail_1H['Evening']=pivot_AzureActiveDirectory_Logfail_1H['18.00PM']
+pivot_AzureActiveDirectory_Logfail_1H['19.00PM']+pivot_AzureActiveDirectory_Logfail_1H['20.00P
M']+pivot_AzureActiveDirectory_Logfail_1H['21.00PM']+pivot_AzureActiveDirectory_Logfail_1H['22.0
0PM']+pivot_AzureActiveDirectory_Logfail_1H['23.00PM']
pivot_AzureActiveDirectory_Logfail_1H['Night']=pivot_AzureActiveDirectory_Logfail_1H['00.00AM']+
pivot_AzureActiveDirectory_Logfail_1H['01.00AM']+pivot_AzureActiveDirectory_Logfail_1H['02.00AM
']+pivot_AzureActiveDirectory_Logfail_1H['03.00AM']+pivot_AzureActiveDirectory_Logfail_1H['04.00
AM']+pivot_AzureActiveDirectory_Logfail_1H['05.00AM']
#Select Column to Show
pivot_AzureActiveDirectory_Logfail_1H=pivot_AzureActiveDirectory_Logfail_1H[['Morning','Afternoo
n','Evening','Night']]
pivot_AzureActiveDirectory_Logfail_1H
#reset index
rein_pivot_AzureActiveDirectory_Logfail_1H=pivot_AzureActiveDirectory_Logfail_1H.reset_index()
rein_pivot_AzureActiveDirectory_Logfail_1H
```

ภาคผนวก ก.3 ชุดคำสั่งไพทอนในการเตรียมตรวจจับสิ่งผิดปรกติ วิธี LOF เข้าใช้ระบบที่มีสถานะสำเร็จ

```
# Importing the libraries
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.neighbors import LocalOutlierFactor
```

```
from sklearn import preprocessing #ทำ Normalization
df_userlogin=pd.read_csv('ตำแหน่งที่เก็บไฟล์')
df_userloginX1 = df_userlogin.iloc[:,[2,3,4,5]]
X1_Normalized = preprocessing.StandardScaler()
X1_Std = X1_Normalized.fit_transform(X1)
#การรันวิธีLocal Outlier Factor
ModelLOF = LocalOutlierFactor(n_neighbors=20)
y_pred1 = ModelLOF.fit_predict(X1_Std) #predict
LOF1 = ModelLOF.negative_outlier_factor_
LOF1 = pd.DataFrame(LOF1)
LOF1.columns = ["LOF"]
print(LOF1)
print(y_pred1)
y_pred1 = pd.DataFrame(y_pred1)
y_pred1["Predict"] = pd.DataFrame(y_pred1)
y_pred1 = y_pred1.drop(y_pred1.columns[[0]], axis=1)
df_Predictuserlogin = pd.concat([df_userlogin,LOF1,y_pred1], axis=1)
df_Predictuserlogin
df_Noruserlogin = df_Predictuserlogin.loc[df_Predictuserlogin['Predict']==1]
df_Outuserlogin = df_Predictuserlogin.loc[df_Predictuserlogin['Predict']==-1]
df_Outuserlogin
pct_out_UserLogin = len(df_Outuserlogin)/len(df_Predictuserlogin) *100
print("pct_out:",pct_out_UserLogin)
print("count Outlier:",len(df_Outuserlogin))
```

ภาคผนวก ก.4 ชุดคำสั่งไพทอนในการเตรียมตรวจจับสิ่งผิดปรกติ วิธี LOF เข้าใช้ระบบที่มีสถานะล้มเหลว

```
# Importing the libraries
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.neighbors import LocalOutlierFactor
from sklearn import preprocessing #ทำ Normalization
df_userlogfail=pd.read_csv('ตำแหน่งที่เก็บไฟล์')
df_userlogfailX1 = df_userlogfail.iloc[:,[2,3,4,5]]
X1_Normalized = preprocessing.StandardScaler()
X1_Std = X1_Normalized.fit_transform(X1)
#การรันวิธีLocal Outlier Factor
ModelLOF = LocalOutlierFactor(n_neighbors=20)
y_pred1 = ModelLOF.fit_predict(X1_Std) #predict
LOF1 = ModelLOF.negative_outlier_factor_
LOF1 = pd.DataFrame(LOF1)
LOF1.columns = ["LOF"]
print(LOF1)
```

```
print(y_pred1)
y_pred1 = pd.DataFrame(y_pred1)
y_pred1["Predict"] = pd.DataFrame(y_pred1)
y_pred1 = y_pred1.drop(y_pred1.columns[[0]], axis=1)
df_Predictuserlogfail = pd.concat([df_userlogfail,LOF1,y_pred1], axis=1)
df_Predictuserlogfail
df_Noruserlogfail = df_Predictuserlogfail.loc[df_Predictuserlogfail['Predict']==1]
df_Outuserlogfail = df_Predictuserlogfail.loc[df_Predictuserlogfail['Predict']==-1]
df_Outuserlogfail
pct_out_UserLogfail = len(df_Outuserlogfail)/len(df_Predictuserlogfail) *100
print("pct_out:",pct_out_UserLogfail)
print("count Outlier:",len(df_Outuserlogfail))
```

ภาคผนวก ก.5 ชุดคำสั่งไพทอนในการเตรียมตรวจจับสิ่งผิดปรกติ วิธี OCSVM เข้าใช้ระบบที่มีสถานะสำเร็จ

```
# Importing the libraries
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.svm import OneClassSVM # import model OCSVM
from sklearn import preprocessing #ทำ Normalization
df_userlogin=pd.read_csv('ตำแหน่งที่เก็บไฟล์')
X1 = df_userlogin.iloc[:,[2,3,4,5]]
X1_Normalized = preprocessing.StandardScaler()
X1_Std = X1_Normalized.fit_transform(X1)
X1_Std
#การรันวิธี One-Class Support Vector Machine
Modelocsvm= OneClassSVM(nu=0.02,kernel="rbf",gamma=0.1)
Modelocsvm.fit(X1_Std)
y_pred1 = Modelocsvm.predict(X1_Std)
SVM1 = Modelocsvm.score_samples(X1_Std) # doctest: +ELLIPSIS
SVM1 = pd.DataFrame(SVM1)
SVM1.columns = ["SVM"]
print(y_pred1)
y_pred1 = pd.DataFrame(y_pred1)
y_pred1["Predict"] = pd.DataFrame(y_pred1)
y_pred1 = y_pred1.drop(y_pred1.columns[[0]], axis=1)
df_Predictuserlogin = pd.concat([df_userlogin,SVM1,y_pred1], axis=1)
df_Predictuserlogin
df_Noruserlogin = df_Predictuserlogin.loc[df_Predictuserlogin['Predict']==1]
df_Outuserlogin = df_Predictuserlogin.loc[df_Predictuserlogin['Predict']==-1]
df_Outuserlogin
pct_out_UserLogin = len(df_Outuserlogin)/len(df_Predictuserlogin) *100
print("pct_out:",pct_out_UserLogin)
```

```
print("count Outlier:",len(df_Outuserlogin))
```

ภาคผนวก ก.6 ชุดคำสั่งไพทอนในการเตรียมตรวจจับสิ่งผิดปรกติ วิธี OCSVM เข้าใช้ระบบที่มีสถานะล้มเหลว

```
# Importing the libraries
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.svm import OneClassSVM # import model OCSVM
from sklearn import preprocessing #ทำ Normalization
df_userlogfail=pd.read_csv('ตำแหน่งที่เก็บไฟล์')
X1 = df_userlogfail.iloc[:,[2,3,4,5]]
X1_Normalized = preprocessing.StandardScaler()
X1_Std = X1_Normalized.fit_transform(X1)
X1_Std
#การรันวิธี One-Class Support Vector Machine
Modelocsvm= OneClassSVM(nu=0.02,kernel="rbf",gamma=0.1)
Modelocsvm.fit(X1_Std)
y_pred1 = Modelocsvm.predict(X1_Std)
SVM1 = Modelocsvm.score_samples(X1_Std) # doctest: +ELLIPSIS
SVM1 = pd.DataFrame(SVM1)
SVM1.columns = ["SVM"]
print(y_pred1)
y_pred1 = pd.DataFrame(y_pred1)
y_pred1["Predict"] = pd.DataFrame(y_pred1)
y_pred1 = y_pred1.drop(y_pred1.columns[[0]], axis=1)
df_Predictuserlogfail = pd.concat([df_userlogfail,SVM1,y_pred1], axis=1)
df_Predictuserlogfail
df_Noruserlogfail = df_Predictuserlogfail.loc[df_Predictuserlogfail['Predict']==1]
df_Outuserlogfail = df_Predictuserlogfail.loc[df_Predictuserlogifail['Predict']==-1]
df_Outuserlogfail
pct_out_UserLogfail = len(df_Outuserlogfail)/len(df_Predictuserlogfail) *100
print("pct_out:",pct_out_UserLogfail)
print("count Outlier:",len(df_Outuserlogfail))
```

ภาคผนวก ก.7 ชุดคำสั่งไพทอนในการเตรียมตรวจจับสิ่งผิดปรกติ วิธี Isolation Forest เข้าใช้ระบบที่มีสถานะสำเร็จ

```
# Importing the libraries
import matplotlib.pyplot as plt
import pandas as pd
from sklearn import preprocessing #ทำ Normalization
from sklearn.ensemble import IsolationForest
df_userlogin=pd.read_csv('ตำแหน่งที่เก็บไฟล์')
```

```
df_userlogin
X1 = df_userlogin.iloc[:,[2,3,4,5]]
X1_Normalized = preprocessing.StandardScaler()
X1_Std = X1_Normalized.fit_transform(X1)
X1_Std
X_scaled_df_login = pd.DataFrame(X1_Std)
X_scaled_df_login
#สร้างแบบจำลองป่าไม้โดดเดี่ยว
IF_UserLogin=
IsolationForest(n_estimators=1000,max_samples='auto',contamination=0.02,max_features=1.0)
IF_UserLogin.fit(X_scaled_df_login)
predictions_UserLogin= IF_UserLogin.fit_predict(X_scaled_df_login)
predictions_UserLogin
# Calculate anomaly scores for each observation
scores = IF_UserLogin.score_samples(X_scaled_df_login)
scores=pd.DataFrame(scores)
scores
concat_UseLogin = pd.concat([df_userlogin,scores], axis=1)
concat_UseLogin.rename(columns={0: "scores"},inplace=True)
concat_UseLogin
y_pred_UseLogin= pd.DataFrame(predictions_UserLogin)
y_pred_UseLogin
concat_UseLogin = pd.concat([concat_UseLogin,y_pred_UseLogin], axis=1)
concat_UseLogin.rename(columns={0: "Predict"},inplace=True)
concat_UseLogin
df_nor_UseLogin = concat_UseLogin.loc[concat_UseLogin['Predict']==1]
df_out_UseLogin = concat_UseLogin.loc[concat_UseLogin['Predict']==-1]
df_out_UseLogin
pct_out_UserLogin = len(df_out_UseLogin)/len(concat_UseLogin) *100
print("pct_out:",pct_out_UserLogin)
print("count Outlier:",len(df_out_UseLogin))
```

ภาคผนวก ก.8 ชุดคำสั่งไพทอนในการเตรียมตรวจจับสิ่งผิดปรกติ วิธี Isolation Forest เข้าใช้ระบบที่มีสถานะล้มเหลว

```
# Importing the libraries
import matplotlib.pyplot as plt
import pandas as pd
from sklearn import preprocessing #ทำ Normalization
from sklearn.ensemble import IsolationForest
df_userlogfail=pd.read_csv('ตำแหน่งที่เก็บไฟล์')
df_userlogfail
X1 = df_userlogfail.iloc[:,[2,3,4,5]]
```

```
X1_Normalized = preprocessing.StandardScaler()
X1_Std = X1_Normalized.fit_transform(X1)
X1_Std
X_scaled_df_logfail = pd.DataFrame(X1_Std)
X_scaled_df_logfail
#สร้างแบบจำลองป่าไม้โดดเดี่ยว
IF_UserLogFail=
IsolationForest(n_estimators=1000,max_samples='auto',contamination=0.02,max_features=1.0)
IF_UserLogFail.fit(X_scaled_df_logfail)
predictions_UserLogfail= IF_UserLogFail.fit_predict(X_scaled_df_logfail)
predictions_UserLogfail
# Calculate anomaly scores for each observation
scores = IF_UserLogFail.score_samples(X_scaled_df_logfail)
scores=pd.DataFrame(scores)
scores
concat_UseLogfail = pd.concat([df_userlogfail,scores], axis=1)
concat_UseLogfail.rename(columns={0: "scores"},inplace=True)
concat_UseLogfail
y_pred_UseLogfail= pd.DataFrame(predictions_UserLogfail)
y_pred_UseLogfail
concat_UseLogfail = pd.concat([concat_UseLogfail,y_pred_UseLogfail], axis=1)
concat_UseLogfail.rename(columns={0: "Predict"},inplace=True)
concat_UseLogfail
df_nor_UseLogfail = concat_UseLogfail.loc[concat_UseLogfail['Predict']==1]
df_out_UseLogfail = concat_UseLogfail.loc[concat_UseLogfail['Predict']==-1]
df_out_UseLogfail
pct_out_UserLogfail = len(df_out_UseLogfail)/len(concat_UseLogfail) *100
print("pct_out:",pct_out_UserLogfail)
print("count Outlier:",len(df_out_UseLogfail))
```

ภาคผนวก ก.9 ชุดคำสั่งไพทอนในการเตรียมตรวจจับสิ่งผิดปรกติ วิธี IF-LOF เข้าใช้ระบบที่มีสถานะสำเร็จ

```
# Importing the libraries
import matplotlib.pyplot as plt
import pandas as pd
from sklearn import preprocessing #ทำ Normalization
from sklearn.ensemble import IsolationForest
from sklearn.neighbors import LocalOutlierFactor
df_userlogin=pd.read_csv('ตำแหน่งที่เก็บไฟล์')
df_userlogin
X1 = df_userlogin.iloc[:,[2,3,4,5]]
X1_Normalized = preprocessing.StandardScaler()
```

```
X1_Std = X1_Normalized.fit_transform(X1)
X1_Std
X_scaled_df_login = pd.DataFrame(X1_Std)
X_scaled_df_login
#สร้างแบบจำลองป่าไม้โดดเดี่ยว
IF_UserLogin=
IsolationForest(n_estimators=1000,max_samples='auto',contamination=0.02,max_features=1.0)
IF_UserLogin.fit(X_scaled_df_login)
predictions_UserLogin= IF_UserLogin.fit_predict(X_scaled_df_login)
predictions_UserLogin
# Calculate anomaly scores for each observation
scores = IF_UserLogin.score_samples(X_scaled_df_login)
scores=pd.DataFrame(scores)
scores
concat_UseLogin = pd.concat([df_userlogin,scores], axis=1)
concat_UseLogin.rename(columns={0: "scores"},inplace=True)
concat_UseLogin
y_pred_UseLogin= pd.DataFrame(predictions_UserLogin)
y_pred_UseLogin
concat_UseLogin = pd.concat([concat_UseLogin,y_pred_UseLogin], axis=1)
concat_UseLogin.rename(columns={0: "Predict"},inplace=True)
concat_UseLogin
df_nor_UseLogin = concat_UseLogin.loc[concat_UseLogin['Predict']==1]
df_out_UseLogin = concat_UseLogin.loc[concat_UseLogin['Predict']==-1]
df_out_UseLogin
pct_out_UserLogin = len(df_out_UseLogin)/len(concat_UseLogin) *100
print("pct_out:",pct_out_UserLogin)
print("count Outlier:",len(df_out_UseLogin))
#การรันวิธีLocal Outlier Factor
ModelLOF = LocalOutlierFactor(n_neighbors=20,contamination=0.02)
y_pred1 = ModelLOF.fit_predict(df_out_UseLogin) #predict
print(LOF1)
print(y_pred1)
y_pred1 = pd.DataFrame(y_pred1)
y_pred1["PredictIFLOF"] = pd.DataFrame(y_pred1)
y_pred1 = y_pred1.drop(y_pred1.columns[[0]], axis=1)
y_pred1
concat_UseLogin = pd.concat([df_OUTIFlogin,y_pred1], axis=1)
concat_UseLogin
df_Noruserlogin = concat_UseLogin.loc[concat_UseLogin['PredictIFLOF']==1]
df_Outuserlogin = concat_UseLogin.loc[concat_UseLogin['PredictIFLOF']==-1]
df_Outuserlogin
pct_out_UserLogin = len(df_Outuserlogin)/len(df_OUTIFlogin) *100
print("pct_out:",pct_out_UserLogin)
```

```
print("count Outlier:",len(df_Outuserlogin))
```

ภาคผนวก ก.10 ชุดคำสั่งไพทอนในการเตรียมตรวจจับสิ่งผิดปรกติ วิธี IF-LOF เข้าใช้ระบบที่มีสถานะล้มเหลว

```
# Importing the libraries
import matplotlib.pyplot as plt
import pandas as pd
from sklearn import preprocessing #ทำ Normalization
from sklearn.ensemble import IsolationForest
from sklearn.neighbors import LocalOutlierFactor
df_userlogfail=pd.read_csv('ตำแหน่งที่เก็บไฟล์')
df_userlogfail
X1 = df_userlogfail.iloc[:,[2,3,4,5]]
X1_Normalized = preprocessing.StandardScaler()
X1_Std = X1_Normalized.fit_transform(X1)
X1_Std
X_scaled_df_logfail = pd.DataFrame(X1_Std)
X_scaled_df_logfail
#สร้างแบบจำลองป่าไม้โดดเดี่ยว
IF_UserLogfail=
IsolationForest(n_estimators=1000,max_samples='auto',contamination=0.02,max_features=1.0)
IF_UserLogfail.fit(X_scaled_df_login)
predictions_UserLogfail= IF_UserLogin.fit_predict(X_scaled_df_login)
predictions_UserLogfail
# Calculate anomaly scores for each observation
scores = IF_UserLogfail.score_samples(X_scaled_df_login)
scores=pd.DataFrame(scores)
scores
concat_UseLogfail = pd.concat([df_userlogin,scores], axis=1)
concat_UseLogfail.rename(columns={0: "scores"},inplace=True)
concat_UseLogfail
y_pred_UseLogfail= pd.DataFrame(predictions_UserLogfail)
y_pred_UseLogfail
concat_UseLogfail = pd.concat([concat_UseLogfail,y_pred_UseLogfail], axis=1)
concat_UseLogfail.rename(columns={0: "Predict"},inplace=True)
concat_UseLogfail
df_nor_UseLogfail = concat_UseLogfail.loc[concat_UseLogfail['Predict']==1]
df_out_UseLogfail = concat_UseLogfail.loc[concat_UseLogfail['Predict']==-1]
df_out_UseLogfail
pct_out_UserLogifail = len(df_out_UseLogfail)/len(concat_UseLogfail) *100
print("pct_out:",pct_out_UserLogfail)
print("count Outlier:",len(df_out_UseLogfail))
#การรันวิธีLocal Outlier Factor
```

```
ModelLOF = LocalOutlierFactor(n_neighbors=20,contamination=0.02)
y_pred1 = ModelLOF.fit_predict(df_out_UseLogfail) #predict
print(LOF1)
print(y_pred1)
y_pred1 = pd.DataFrame(y_pred1)
y_pred1["PredictIFLOF"] = pd.DataFrame(y_pred1)
y_pred1 = y_pred1.drop(y_pred1.columns[[0]], axis=1)
y_pred1
concat_UseLogfail = pd.concat([df_OUTIFlogin,y_pred1], axis=1)
concat_UseLogfail
df_Noruserlogfail = concat_UseLogfail.loc[concat_UseLogfail['PredictIFLOF']==1]
df_Outuserlogfail = concat_UseLogfail.loc[concat_UseLogfail['PredictIFLOF']==-1]
df_Outuserlogfail
pct_out_UserLogfail = len(df_Outuserlogfail)/len(df_OUTIFlogfail) *100
print("pct_out:",pct_out_UserLogfail)
print("count Outlier:",len(df_Outuserlogfail))
```

ภาคผนวก ก.11 ชุดคำสั่งไพทอนในการสร้างกราฟเพื่อดูข้อมูล

```
#การ plot กราฟ 3 มิติ
# Importing the libraries
from mpl_toolkits.mplot3d import Axes3D
fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')
# Plot x's for the ground truth normal
ax.scatter(df_Noruserlogin.iloc[:, 2], df_Noruserlogin.iloc[:, 4], zs=df_Noruserlogin.iloc[:, 5], s=4,
 lw=0,label="normal")
# Plot x's for the ground truth outliers
ax.scatter(df_Outuserlogin.iloc[:, 2], df_Outuserlogin.iloc[:, 4], zs=df_Outuserlogin.iloc[:, 5],
lw=2, s=50, marker="x", c="red", label="outliers")
ax.set_title('Local Outlier Factor ')
ax.set_xlabel('Morning')
ax.set_ylabel('Afternoon')
ax.set_zlabel('Evening')
ax.legend()
```

ภาคผนวก ก.12 ชุดคำสั่งไพทอนในการจัดกลุ่มด้วยวิธี K-Mean

```
# K-Means Clustering
# Importing the libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.metrics import silhouette_score
```

```
from sklearn.cluster import KMeans
from sklearn.datasets import make_blobs
dataLogin = pd.read_csv('ตำแหน่งที่เก็บไฟล์')
X_in = dataLogin.iloc[:,2:6].values
X_in
# Using the elbow method to find the optimal number of clusters
from sklearn.cluster import KMeans
wcss =[]
for i in range (1,11):
    kmeans = KMeans(n_clusters = i, init = 'k-means++', max_iter =300, n_init = 10, random_state =
0)
    kmeans.fit(X_in)
    wcss.append(kmeans.inertia_)
# Plot the graph to visualize the Elbow Method to find the optimal number of cluster
plt.plot(range(1,11),wcss,marker='o')
plt.title('The Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
# Applying KMeans to the dataset with the optimal number of cluster
kmeans=KMeans(n_clusters= 3, init = 'k-means++', max_iter = 300, n_init = 10, random_state = 0)
y_kmeans = kmeans.fit_predict(X_in)
# calculate the silhouette coefficient
score = silhouette_score(X_in, kmeans.labels_)
print("Silhouette Coefficient:", score)
# Visualising the clusters
plt.scatter(X_in[y_kmeans == 0, 0], X_in[y_kmeans == 0,1],s = 100, c='red', label = 'Cluster 1')
plt.scatter(X_in[y_kmeans == 1, 0], X_in[y_kmeans == 1,1],s = 100, c='blue', label = 'Cluster 2')
plt.scatter(X_in[y_kmeans == 2, 0], X_in[y_kmeans == 2,1],s = 100, c='green', label = 'Cluster 3')
plt.scatter(X_in[y_kmeans == 3, 0], X_in[y_kmeans == 3,1],s = 100, c='cyan', label = 'Cluster 4')
plt.scatter(X_in[y_kmeans == 4, 0], X_in[y_kmeans == 4,1],s = 100, c='magenta', label = 'Cluster 5')
plt.scatter(kmeans.cluster_centers_[:,0], kmeans.cluster_centers_[:,1], s = 300, c = 'yellow', label =
'Centroids')
plt.title('Clusters of user')
plt.xlabel('time')
plt.ylabel('count')
plt.legend()
plt.show()
y_kmeans = pd.DataFrame(y_kmeans)
y_kmeans
con_in = pd.concat([dataLogin,y_kmeans], axis=1)
con_in.rename(columns={0: "cluster"},inplace=True)
con_in
```

ภาคผนวก ข

ข.1 ผลการวิเคราะห์วิธีการตรวจจับสิ่งผิดปรกติ ทั้ง 4 วิธี

ข.1.1 ข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จ

ตารางที่ ข.1 ตัวอย่างวิธีการตรวจจับสิ่งผิดปรกติด้วยวิธี LOF พิจารณาตามช่วงเวลา

	date	user_email	Morning	Afternoon	Evening	Night	predict	LOFScore
1	20/01/2023	AAAAAA	5	5	2	0	1	-0.9820
2	17/01/2023	BBBBBB	9	6	4	0	1	-0.9869
3	25/01/2023	CCCCCC	8	3	2	0	1	-0.9872
4	04/03/2023	TTTTTT	3	7	3	0	1	-0.9503
5	01/03/2023	ZZZZZZ	8	3	2	0	1	-0.9872
...	...	...	...	...	...	...	...	...
115,744	18/01/2023	BBBBBB	198	0	0	328	-1	-2.3374
115,745	02/03/2023	EEEEEE	192	0	0	0	-1	-1.6160
115,746	21/02/2023	HHHHHH	2,229	892	0	283	-1	-8.4339
115,747	16/02/2023	FFFFFF	1,438	1,224	0	1,132	-1	-8.4429
115,748	14/02/2023	GGGGGG	1,423	1,454	1,457	985	-1	-10.3449

จากตารางที่ ข.1 แสดงรายละเอียดของตัวอย่างวิธีการตรวจจับสิ่งผิดปรกติของวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ โดยกำหนดจำนวนเพื่อนบ้านใกล้เคียงเป็น 20 พิจารณาจากช่วงเวลาพบว่าวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ ให้จำนวนค่าที่ผิดปรกติ 1,771 รายการ จาก 115,748 รายการ คิดเป็น 1.53% ของจำนวนรายการทั้งหมดโดยพิจารณาจากค่า LOFScore ที่น้อยกว่า -1.5

ตารางที่ ข.2 ตัวอย่างวิธีการตรวจจับสิ่งผิดปรกติด้วยวิธี OCSVM พิจารณาตามช่วงเวลา

	date	user_email	Morning	Afternoon	Evening	Night	predict	SVMscores
1	09/01/2023	AAAAAA	1	8	0	0	1	317.1094
2	10/01/2023	BBBBBB	4	5	0	0	1	322.9801
3	24/02/2023	CCCCCC	6	8	1	0	1	338.3050
4	04/03/2023	TTTTTT	3	4	3	0	1	335.6952
5	08/03/2023	ZZZZZZ	5	3	0	0	1	320.3808
...	...	...	...	...	...	...	...	...
115,744	17/01/2023	BBBBBB	1411	1260	0	0	-1	1
115,745	16/02/2023	EEEEEE	895	1431	1417	0	-1	1
115,746	21/02/2023	HHHHHH	27	2124	1253	0	-1	1
115,747	03/03/2023	FFFFFF	737	727	423	59	-1	1
115,748	26/01/2023	GGGGGG	0	49	1108	0	-1	1

จากตารางที่ ข.2 แสดงรายละเอียดของตัวอย่างวิธีการตรวจจับสิ่งผิดปรกติของวิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง โดยกำหนดค่า nu (υ) เท่ากับ 0.02 ค่าแกมมาเท่ากับ auto และเคอร์เนล (Kernel) ที่ใช้คือ Radial Basis Function Kernel (RBF) พิจารณาจากช่วงเวลา พบว่า วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่งให้จำนวนค่าที่ผิดปรกติ 2,160 รายการ จาก 115,748 รายการ คิดเป็น 1.86 % ของจำนวนรายการทั้งหมดโดยพิจารณาจากค่า SVM ที่อยู่ในช่วง 1 ถึง 298.2791

ตารางที่ ข.3 ตัวอย่างวิธีการตรวจจับสิ่งผิดปรกติด้วยวิธี Isolation Forest พิจารณาตามช่วงเวลา

	date	user_email	Morning	Afternoon	Evening	Night	predict	IFScores
1	09/01/2023	AAAAAA	7	7	1	0	1	-0.4421
2	18/01/2023	BBBBBB	4	5	2	0	1	-0.4304
3	27/01/2023	CCCCCC	5	4	0	0	1	-0.3550
4	02/02/2023	TTTTTT	7	5	0	0	1	-0.3710
5	20/02/2023	ZZZZZZ	3	5	0	0	1	-0.3519
...	...	...	...	...	...	...	...	...
115,744	16/02/2023	BBBBBB	895	1431	1417	0	-1	-0.8528
115,745	14/02/2023	EEEEEE	737	1436	1449	15	-1	-0.8773
115,746	03/03/2023	HHHHHH	737	727	423	59	-1	-0.8793
115,747	15/02/2023	FFFFFF	1466	51	0	1445	-1	-0.8333
115,748	17/01/2023	GGGGGG	1411	1260	0	0	-1	-0.8011

จากตารางที่ ข.3 แสดงรายละเอียดของตัวอย่างวิธีการตรวจจับสิ่งผิดปรกติของวิธีไอโซเลชั่นฟอเรส โดยกำหนดจำนวนของ Decision Trees = 1,000 จำนวนของ max_samples='auto' จำนวนของ max_features= 4 พบว่า วิธีไอโซเลชั่นฟอเรสให้จำนวนค่าที่ผิดปรกติ 2,315 รายการ จาก 115,748 รายการ คิดเป็น 2.00% ของจำนวนรายการทั้งหมด ซึ่งคะแนนความความผิดปรกตินั้นจะมีค่าน้อยกว่า -0.5 จะหมายถึงข้อมูลที่มีแนวโน้มผิดปรกติ

ตารางที่ ข.4 ตัวอย่างวิธีการตรวจจับสิ่งผิดปรกติด้วยวิธี IF-LOF พิจารณาตามช่วงเวลา

	date	user_email	Morning	Afternoon	Evening	Night	predict	IFscore	LOFscore
1	09/01/2023	AAAAAA	7	7	1	0	1	-0.4421	-0.9908
2	18/01/2023	BBBBBB	4	5	2	0	1	-0.4304	-1.0000
3	27/01/2023	CCCCCC	5	4	0	0	1	-0.3550	-1.0000
4	02/02/2023	TTTTTT	7	5	0	0	1	-0.3710	-1.0000
5	20/02/2023	ZZZZZZ	3	5	0	0	1	-0.3519	-1.0000
...	...	...	...	...	...	...	...	...	...
115,744	16/02/2023	BBBBBB	895	1431	1417	0	-1	-0.8528	-11.0127
115,745	14/02/2023	EEEEEE	737	1436	1449	15	-1	-0.8773	-7.0422
115,746	03/03/2023	HHHHHH	737	727	423	59	-1	-0.8793	-8.6007
115,747	15/02/2023	FFFFFF	1466	51	0	1445	-1	-0.8333	-8.5349
115,748	17/01/2023	GGGGGG	1411	1260	0	0	-1	-0.8011	-4.3710

จากตารางที่ ข.4 แสดงรายละเอียดของตัวอย่างวิธีการตรวจจับสิ่งผิดปรกติของวิธีไอเอฟ-แอลโอเอฟพิจารณาจากช่วงเวลา ให้จำนวนค่าที่ผิดปรกติ 228 รายการ จาก 115,748 รายการ คิดเป็น 0.19% ของจำนวนรายการทั้งหมด ซึ่งคะแนนความความผิดปรกตินั้นจะมีค่าน้อยกว่า -0.5 จะหมายถึงข้อมูลที่มีแนวโน้มผิดปรกติ และสามารถพิจารณาจากค่า LOFScore ที่น้อยกว่า -1.5

ข.1.2 ข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลว

ตารางที่ ข.5 ตัวอย่างวิธีการตรวจจับสิ่งผิดปรกติด้วยวิธี LOF พิจารณาตามช่วงเวลา

	date	user_email	Morning	Afternoon	Evening	Night	predict	LOFScore
1	09/01/2023	AAAAAA	4	8	0	0	1	-0.9392
2	09/01/2023	BBBBBB	2	8	6	0	1	-1.0713
3	21/02/2023	CCCCCC	7	4	0	0	1	-0.9901
4	11/03/2023	TTTTTT	0	7	6	0	1	-0.9870
5	13/01/2023	ZZZZZZ	8	9	8	2	1	-0.9852
...	...	...	...	...	...	...	...	...
54,259	11/03/2023	BBBBBB	215	208	0	302	-1	-2.1963
54,260	13/01/2023	EEEEEE	197	176	159	78	-1	-1.5432
54,261	17/01/2023	HHHHHH	669	0	0	0	-1	-7.4433
54,262	02/02/2023	FFFFFF	464	446	598	596	-1	-1.7912
54,263	04/02/2023	GGGGGG	297	0	0	597	-1	-3.3029

จากตารางที่ ข.5 แสดงรายละเอียดของตัวอย่างวิธีการตรวจจับสิ่งผิดปรกติของวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ โดยกำหนดจำนวนเพื่อนบ้านใกล้เคียงเป็น 20 พิจารณาจากช่วงเวลาพบว่าวิธีโลคอลเอาท์ไลเออร์แฟคเตอร์ ให้จำนวนค่าที่ผิดปรกติ 1,070 รายการ จาก 54,263 รายการ คิดเป็น 1.97% ของจำนวนรายการทั้งหมดโดยพิจารณาจากค่า LOFScore ที่น้อยกว่า -1.5

ตารางที่ ข.6 ตัวอย่างวิธีการตรวจจับสิ่งผิดปรกติด้วยวิธี OCSVM พิจารณาตามช่วงเวลา

	date	user_email	Morning	Afternoon	Evening	Night	predict	SVMscores
1	28/02/2023	AAAAAA	8	0	0	0	1	185.8921
2	28/02/2023	BBBBBB	2	1	1	0	1	187.0699
3	01/03/2023	CCCCCC	8	5	0	0	1	197.2853
4	01/03/2023	TTTTTT	5	6	1	0	1	201.3312
5	02/03/2023	ZZZZZZ	0	3	1	0	1	185.8161
...	...	...	...	...	...	...	...	...
54,259	17/01/2023	BBBBBB	669	0	0	0	-1	1
54,260	29/01/2023	EEEEEE	523	486	418	476	-1	1
54,261	09/03/2023	HHHHHH	445	265	292	126	-1	1
54,262	10/03/2023	FFFFFF	255	351	333	317	-1	1
54,263	09/03/2023	GGGGGG	327	169	259	172	-1	1

จากตารางที่ ข.6 แสดงรายละเอียดของตัวอย่างวิธีการตรวจจับสิ่งผิดปรกติของวิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง โดยกำหนดค่า nu (υ) เท่ากับ 0.02 ค่าแกมมาเท่ากับ auto และเคอร์เนล (Kernel) ที่ใช้คือ Radial Basis Function Kernel (RBF) พิจารณาจากช่วงเวลา พบว่า วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่งให้จำนวนค่าที่ผิดปรกติ 1,003 รายการ จาก 54,263 รายการ คิดเป็น 1.84 % ของจำนวนรายการทั้งหมดโดยพิจารณาจากค่า SVM ที่อยู่ในช่วง 1 ถึง 176.6080

ตารางที่ ข.7 ตัวอย่างวิธีการตรวจจับสิ่งผิดปรกติด้วยวิธี Isolation Forest พิจารณาตามช่วงเวลา

	date	user_email	Morning	Afternoon	Evening	Night	predict	IFScores
1	09/01/2023	AAAAAA	0	6	1	0	1	-0.4538
2	09/01/2023	BBBBBB	0	3	2	0	1	-0.4322
3	24/01/2023	CCCCCC	2	1	0	0	1	-0.3484
4	13/02/2023	TTTTTT	1	3	0	0	1	-0.3718
5	27/02/2023	ZZZZZZ	5	3	0	0	1	-0.4347
...	...	...	...	...	...	...	...	...
54,259	30/01/2023	BBBBBB	486	516	536	470	-1	-0.8885
54,260	10/03/2023	EEEEEE	282	518	217	203	-1	-0.8855
54,261	09/03/2023	HHHHHH	327	169	259	172	-1	-0.8842
54,262	10/03/2023	FFFFFF	212	357	232	352	-1	-0.8849
54,263	16/01/2023	GGGGGG	290	300	102	20	-1	-0.8767

จากตารางที่ ข.7 แสดงรายละเอียดของตัวอย่างวิธีการตรวจจับสิ่งผิดปรกติของวิธีไอโซเลชั่นฟอเรส โดยกำหนดจำนวนของ Decision Trees = 1,000 จำนวนของ max_samples='auto' จำนวนของ max_features= 4 พบว่า วิธีไอโซเลชั่นฟอเรสให้จำนวนค่าที่ผิดปรกติ 1,081 รายการ จาก 54,263 รายการ คิดเป็น 1.99% ของจำนวนรายการทั้งหมด ซึ่งคะแนนความความผิดปรกตินั้นจะมีค่าน้อยกว่า -0.5 จะหมายถึงข้อมูลที่มีแนวโน้มผิดปรกติ

ตารางที่ ข.8 ตัวอย่างวิธีการตรวจจับสิ่งผิดปรกติด้วยวิธี IF-LOF พิจารณาตามช่วงเวลา

	date	user_email	Morning	Afternoon	Evening	Night	predict	IFscore	LOFscore
1	09/01/2023	AAAAAA	0	6	1	0	1	-0.4538	-1
2	09/01/2023	BBBBBB	0	3	2	0	1	-0.4322	-0.9999
3	24/01/2023	CCCCCC	2	1	0	0	1	-0.3484	-1
4	13/02/2023	TTTTTT	1	3	0	0	1	-0.3718	-1
5	27/02/2023	ZZZZZZ	5	3	0	0	1	-0.4347	-1
...	...	...	...	...	...	...	...	...	...
54,259	17/01/2023	BBBBBB	669	0	0	0	-1	-0.6908	-6.8636
54,260	11/03/2023	EEEEEE	325	77	0	295	-1	-0.8546	-2.3231
54,261	29/01/2023	HHHHHH	523	486	418	476	-1	-0.8883	-1.5418
54,262	16/01/2023	FFFFFF	290	300	102	20	-1	-0.8767	-1.6361
54,263	13/03/2023	GGGGGG	347	0	0	0	-1	-0.6904	-4.2456

จากตารางที่ ข.8 แสดงรายละเอียดของตัวอย่างวิธีการตรวจจับสิ่งผิดปรกติของวิธีไอเอฟ-แอลโอเอฟ พิจารณาจากช่วงเวลา ให้จำนวนค่าที่ผิดปรกติ 167 รายการ จาก 54,263 รายการ คิดเป็น 0.30 % ของจำนวนรายการทั้งหมด ซึ่งคะแนนความความผิดปรกตินั้นจะมีค่าน้อยกว่า -0.5 จะหมายถึงข้อมูลที่มีแนวโน้มผิดปรกติ และสามารถพิจารณาจากค่า LOFScore ที่น้อยกว่า -1.5

ภาคผนวก ค

ค.1 ผลการวิเคราะห์การแจกแจงปรกติของการจัดกลุ่มเพื่อที่จะนำไปทดสอบสมมติฐาน

ค.1.1 ข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จ

ค.1.1.1 วิธีโลคอลเอาท์ไลเออร์แฟคเตอร์

Tests of Normality

	cluster	Kolmogorov-Smirnov[a] Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
Morning	0	.401	1765	.000	.193	1765	.000
	1	.241	6	.200*	.903	6	.393
Afternoon	0	.369	1765	.000	.245	1765	.000
	1	.265	6	.200*	.937	6	.639
Evening	0	.441	1765	.000	.089	1765	.000
	1	.243	6	.200*	.892	6	.328
Night	0	.461	1765	.000	.063	1765	.000
	1	.366	6	.012	.636	6	.001

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

รูปที่ ค.1 ผลลัพธ์การวิเคราะห์การแจกแจงข้อมูล

จากรูปที่ ค.1 พบว่าข้อมูลส่วนใหญ่ไม่มีการแจกแจงปรกติ จึงเลือกใช้การใช้สถิติแบบไม่ใช้พารามิเตอร์ (Nonparametric Statistics)

ค.1.1.2 วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง

Tests of Normality

	cluster	Kolmogorov-Smirnov[a] Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
Morning	0	.285	2154	.000	.457	2154	.000
	1	.241	6	.200*	.903	6	.393
Afternoon	0	.232	2154	.000	.660	2154	.000
	1	.265	6	.200*	.937	6	.639
Evening	0	.370	2154	.000	.313	2154	.000
	1	.243	6	.200*	.892	6	.328
Night	0	.451	2154	.000	.079	2154	.000
	1	.366	6	.012	.636	6	.001

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

รูปที่ ค.2 ผลลัพธ์การวิเคราะห์การแจกแจงข้อมูล

จากรูปที่ ค.2 พบว่าข้อมูลส่วนใหญ่ไม่มีการแจกแจงปรกติ จึงเลือกใช้การใช้สถิติแบบไม่ใช้พารามิเตอร์ (Nonparametric Statistics)

ค.1.1.3 วิธีไอโซเลชั่นฟอเรส

Tests of Normality

	cluster	Kolmogorov-Smirnov[a] Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
Morning	0	.310	2309	.000	.417	2309	.000
	1	.241	6	.200*	.903	6	.393
Afternoon	0	.242	2309	.000	.636	2309	.000
	1	.265	6	.200*	.937	6	.639
Evening	0	.357	2309	.000	.323	2309	.000
	1	.243	6	.200*	.892	6	.328
Night	0	.444	2309	.000	.085	2309	.000
	1	.366	6	.012	.636	6	.001

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

รูปที่ ค.3 ผลลัพธ์การวิเคราะห์การแจกแจงข้อมูล

จากรูปที่ ค.3 พบว่าข้อมูลส่วนใหญ่ไม่มีการแจกแจงปรกติ จึงเลือกใช้การใช้สถิติแบบไม่ใช้พารามิเตอร์ (Nonparametric Statistics)

ค.1.1.4 วิธีไอเอฟ-แอลโอเอฟ

Tests of Normality

	cluster	Kolmogorov-Smirnov[a] Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
Morning	0	.334	222	.000	.455	222	.000
	1	.241	6	.200*	.903	6	.393
Afternoon	0	.328	222	.000	.477	222	.000
	1	.265	6	.200*	.937	6	.639
Evening	0	.393	222	.000	.245	222	.000
	1	.243	6	.200*	.892	6	.328
Night	0	.397	222	.000	.228	222	.000
	1	.366	6	.012	.636	6	.001

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

รูปที่ ค.4 ผลลัพธ์การวิเคราะห์การแจกแจงข้อมูล

จากรูปที่ ค.4 พบว่าข้อมูลส่วนใหญ่ไม่มีการแจกแจงปรกติ จึงเลือกใช้การใช้สถิติแบบไม่ใช้พารามิเตอร์ (Nonparametric Statistics)

ค.1.2 ข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลว

ค.1.2.1 วิธีโลคอลเอาท์ไลเออร์แฟคเตอร์

Tests of Normality

	cluster	Kolmogorov-Smirnov[a] Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
Morning	0	.380	1064	.000	.302	1064	.000
	1	.158	6	.200*	.952	6	.758
Afternoon	0	.366	1064	.000	.359	1064	.000
	1	.191	6	.200*	.884	6	.286
Evening	0	.392	1064	.000	.277	1064	.000
	1	.241	6	.200*	.957	6	.796
Night	0	.422	1064	.000	.181	1064	.000
	1	.193	6	.200*	.888	6	.308

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

รูปที่ ค.5 ผลลัพธ์การวิเคราะห์การแจกแจงข้อมูล

จากรูปที่ ค.5 พบว่าข้อมูลส่วนใหญ่ไม่มีการแจกแจงปรกติ จึงเลือกใช้การใช้สถิติแบบไม่ใช้พารามิเตอร์ (Nonparametric Statistics)

ค.1.2.2 วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง

Tests of Normality

	cluster	Kolmogorov-Smirnov[a] Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
Morning	0	.256	974	.000	.684	974	.000
	1	.198	29	.005	.933	29	.068
Afternoon	0	.269	974	.000	.648	974	.000
	1	.101	29	.200*	.957	29	.279
Evening	0	.329	974	.000	.486	974	.000
	1	.124	29	.200*	.945	29	.133
Night	0	.357	974	.000	.411	974	.000
	1	.105	29	.200*	.924	29	.040

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

รูปที่ ค.6 ผลลัพธ์การวิเคราะห์การแจกแจงข้อมูล

จากรูปที่ ค.6 พบว่าข้อมูลส่วนใหญ่ไม่มีการแจกแจงปรกติ จึงเลือกใช้การใช้สถิติแบบไม่ใช้พารามิเตอร์ (Nonparametric Statistics)

ค.1.2.3 วิธีไอโซเลชั่นฟอเรส

Tests of Normality

	cluster	Kolmogorov-Smirnov[a] Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
Morning	0	.284	1052	.000	.593	1052	.000
	1	.198	29	.005	.933	29	.068
Afternoon	0	.273	1052	.000	.610	1052	.000
	1	.101	29	.200*	.957	29	.279
Evening	0	.319	1052	.000	.477	1052	.000
	1	.124	29	.200*	.945	29	.133
Night	0	.329	1052	.000	.448	1052	.000
	1	.105	29	.200*	.924	29	.040

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

รูปที่ ค.7 ผลลัพธ์การวิเคราะห์การแจกแจงข้อมูล

จากรูปที่ ค.7 พบว่าข้อมูลส่วนใหญ่ไม่มีการแจกแจงปรกติ จึงเลือกใช้การใช้สถิติแบบไม่ใช้พารามิเตอร์ (Nonparametric Statistics)

ค.1.2.4 วิธีไอเอฟ-แอลโอเอฟ

Tests of Normality

	cluster	Kolmogorov-Smirnov[a] Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
Morning	0	.269	161	.000	.645	161	.000
	1	.158	6	.200*	.952	6	.758
Afternoon	0	.235	161	.000	.753	161	.000
	1	.191	6	.200*	.884	6	.286
Evening	0	.310	161	.000	.658	161	.000
	1	.241	6	.200*	.957	6	.796
Night	0	.321	161	.000	.497	161	.000
	1	.193	6	.200*	.888	6	.308

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

รูปที่ ค.8 ผลลัพธ์การวิเคราะห์การแจกแจงข้อมูล

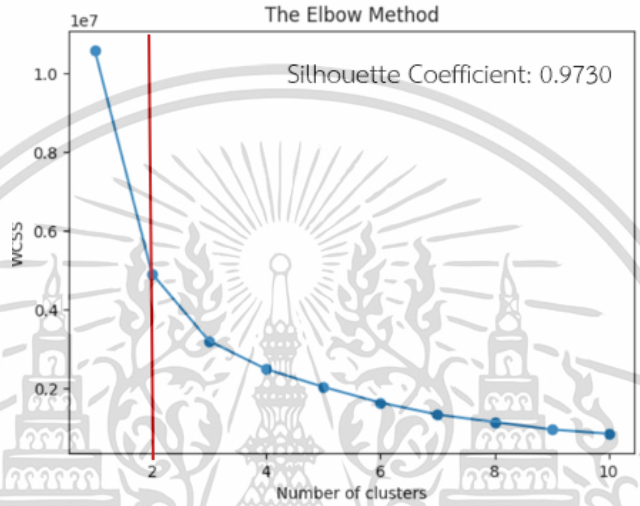
จากรูปที่ ค.8 พบว่าข้อมูลส่วนใหญ่ไม่มีการแจกแจงปรกติ จึงเลือกใช้การใช้สถิติแบบไม่ใช้พารามิเตอร์ (Nonparametric Statistics)

ภาคผนวก ง

ง.1 รูปการแบ่งกลุ่มของสิ่งผิดปรกติ

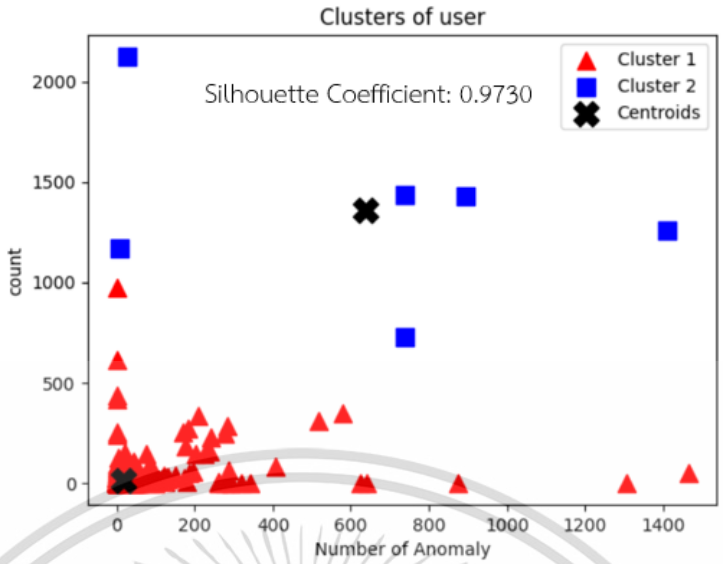
ง.1.1 ข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จ

ง.1.1.1 วิธีโลคอลเอาท์ไลเออร์แฟคเตอร์



รูปที่ ง.1 Elbow หาค่า K ที่เหมาะสม และค่า Silhouette ของวิธี LOF

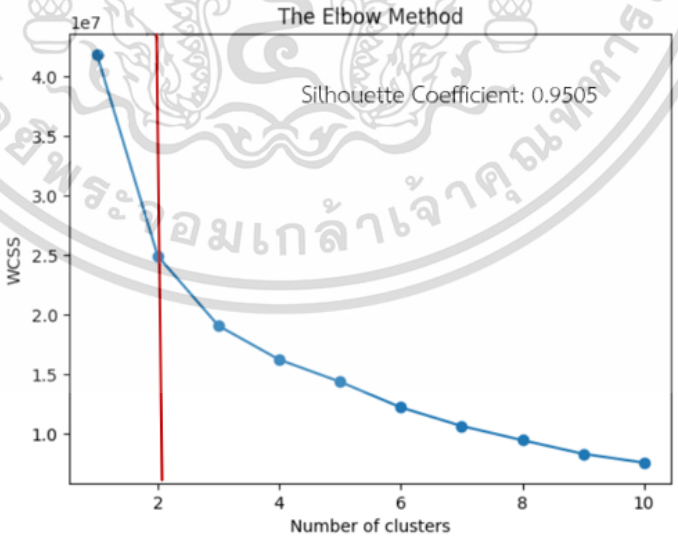
รูปที่ ง.1 จะพบว่าจำนวนกลุ่มที่เหมาะสมในการจัดกลุ่มคือ 2 กลุ่ม (K = 2) และค่า Silhouette Coefficient มีค่าเท่ากับ 0.973 ซึ่งมีค่าใกล้ 1 แสดงว่าข้อมูลถูกจัดให้อยู่ในกลุ่มมีความเหมาะสมแล้ว ซึ่งถ้ามองจากภาพแผนภาพ Elbow ที่มีจุดหักศอกที่ 2 ก็มีความเป็นไปได้ที่ข้อมูลจะแบ่งเป็น 2 กลุ่มเช่นเดียวกัน แสดงว่าการจัดกลุ่ม 2 กลุ่มเหมาะสมกับข้อมูลชุดนี้ที่สุด จึงทำการจัดกลุ่มข้อมูล 2 กลุ่ม



รูปที่ ง.2 รูปการแบ่งกลุ่มของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จของวิธี LOF

จากรูปที่ ง.2 พบว่าข้อมูลส่วนใหญ่ของกลุ่มที่ 1 ข้อมูลมีการเกาะกลุ่มกันและมีการกระจายตัวที่น้อย และกลุ่มที่ 2 ข้อมูลมีการกระจายตัวที่มาก และค่า Silhouette Coefficient มีค่าเท่ากับ 0.973 ซึ่งมีค่าใกล้ 1 แสดงว่าข้อมูลถูกจัดให้อยู่ในกลุ่มมีความเหมาะสมแล้ว

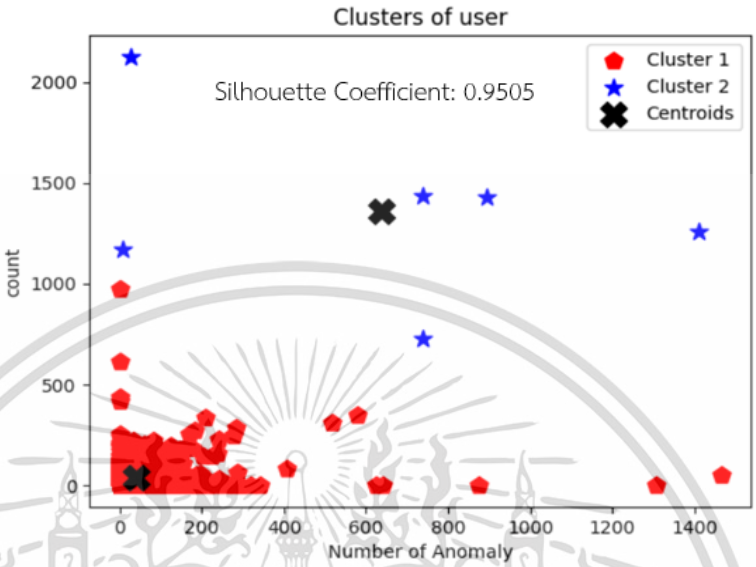
ง.1.1.2 วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง



รูปที่ ง.3 Elbow หาค่า K ที่เหมาะสม และค่า Silhouetteของวิธี OCSVM

รูปที่ ง.3 จะพบว่าจำนวนกลุ่มที่เหมาะสมในการจัดกลุ่มคือ 2 กลุ่ม (K = 2) และค่า Silhouette Coefficient มีค่าเท่ากับ 0.9505 ซึ่งมีค่าใกล้ 1 แสดงว่าข้อมูลถูกจัดให้อยู่ในกลุ่มมีความ

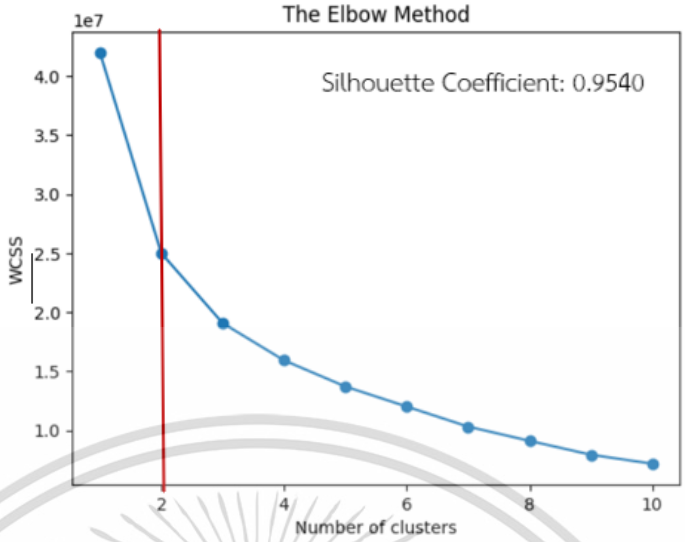
เหมาะสมแล้ว ซึ่งถ้ามองจากภาพแผนภาพ Elbow ที่มีจุดหักศอกที่ 2 ก็มีความเป็นไปได้ที่ข้อมูลจะแบ่งเป็น 2 กลุ่มเช่นเดียวกัน แสดงว่าการจัดกลุ่ม 2 กลุ่มเหมาะสมกับข้อมูลชุดนี้ที่สุด จึงทำการจัดกลุ่มข้อมูล 2 กลุ่ม



รูปที่ ง.4 รูปการแบ่งกลุ่มของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จของวิธี OCSVM

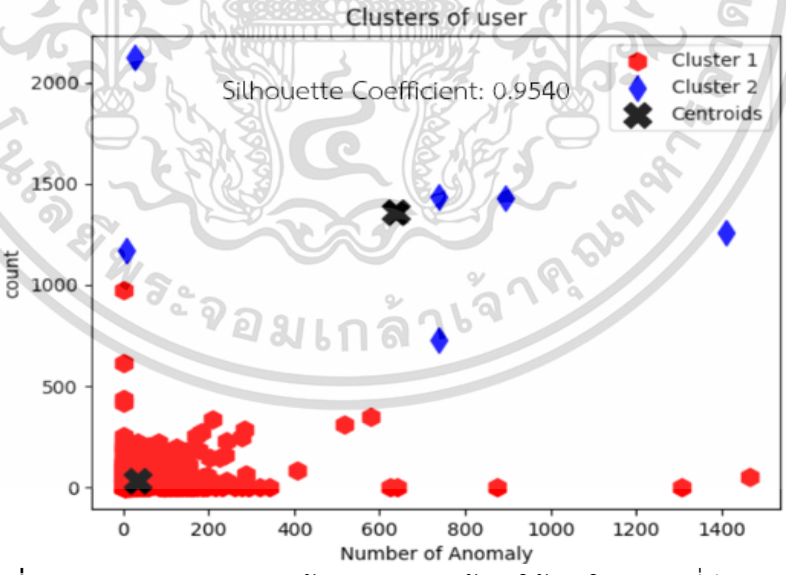
จากรูปที่ ง.4 พบว่าข้อมูลส่วนใหญ่ของกลุ่มที่ 1 ข้อมูลมีการเกาะกลุ่มกันและมีการกระจายตัวที่น้อย และกลุ่มที่ 2 ข้อมูลมีการกระจายตัวที่มาก และค่า Silhouette Coefficient มีค่าเท่ากับ 0.9505 ซึ่งมีค่าใกล้ 1 แสดงว่าข้อมูลถูกจัดให้อยู่ในกลุ่มมีความเหมาะสมแล้ว

ง.1.1.3 วิธีไอโซเลชั่นฟอเรส



รูปที่ ง.5 Elbow หาค่า K ที่เหมาะสม และค่า Silhouette ของวิธี Isolation Forest

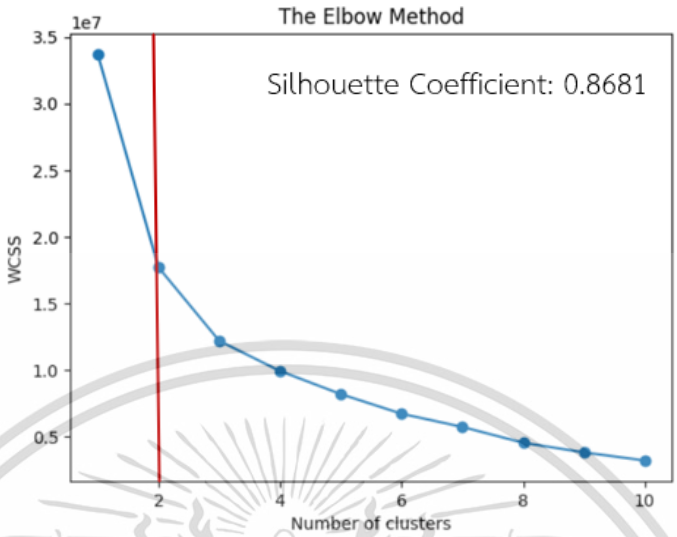
รูปที่ ง.5 จะพบว่าจำนวนกลุ่มที่เหมาะสมในการจัดกลุ่มคือ 2 กลุ่ม (K = 2) และค่า Silhouette Coefficient มีค่าเท่ากับ 0.9540 ซึ่งมีค่าใกล้ 1 แสดงว่าข้อมูลถูกจัดให้อยู่ในกลุ่มมีความเหมาะสมแล้ว ซึ่งถ้ามองจากภาพแผนภาพ Elbow ที่มีจุดหักศอกที่ 2 ก็มีความเป็นไปได้ที่ข้อมูลจะแบ่งเป็น 2 กลุ่มเช่นเดียวกัน แสดงว่าการจัดกลุ่ม 2 กลุ่มเหมาะสมกับข้อมูลชุดนี้ที่สุด จึงทำการจัดกลุ่มข้อมูล 2 กลุ่ม



รูปที่ ง.6 รูปการแบ่งกลุ่มของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จของวิธี Isolation Forest

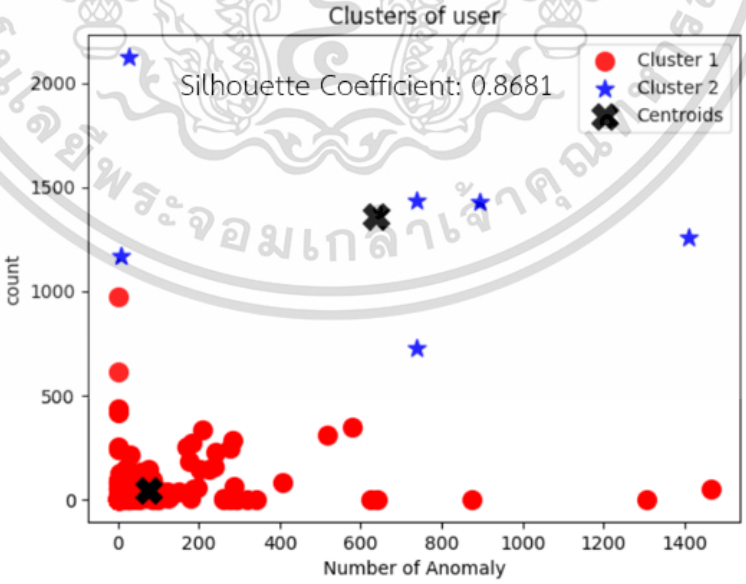
จากรูปที่ ง.6 พบว่าข้อมูลส่วนใหญ่ของกลุ่มที่ 1 ข้อมูลมีการเกาะกลุ่มกันและมีการกระจายตัวที่น้อย และกลุ่มที่ 2 ข้อมูลมีการกระจายตัวที่มาก และค่า Silhouette Coefficient มีค่าเท่ากับ 0.9540 ซึ่งมีค่าใกล้ 1 แสดงว่าข้อมูลถูกจัดให้อยู่ในกลุ่มมีความเหมาะสมแล้ว

ง.1.1.4 วิธีไอเอฟ-แอลโอเอฟ



รูปที่ ง.7 Elbow หาค่า K ที่เหมาะสม และค่า Silhouette ของวิธี IF-LOF

รูปที่ ง.7 จะพบว่าจำนวนกลุ่มที่เหมาะสมในการจัดกลุ่มคือ 2 กลุ่ม (K = 2) และค่า Silhouette Coefficient มีค่าเท่ากับ 0.8681 ซึ่งมีค่าใกล้ 1 แสดงว่าข้อมูลถูกจัดให้อยู่ในกลุ่มมีความเหมาะสมแล้ว ซึ่งถ้ามองจากภาพแผนภาพ Elbow ที่มีจุดหักศอกที่ 2 ก็มีความเป็นไปได้ที่ข้อมูลจะแบ่งเป็น 2 กลุ่มเช่นเดียวกัน แสดงว่าการจัดกลุ่ม 2 กลุ่มเหมาะสมกับข้อมูลชุดนี้ที่สุด จึงทำการจัดกลุ่มข้อมูล 2 กลุ่ม

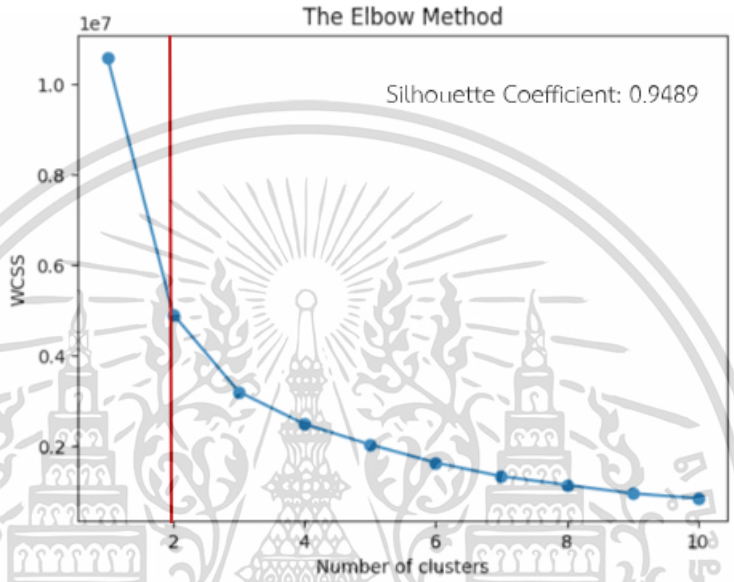


รูปที่ ง.8 รูปการแบ่งกลุ่มของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะสำเร็จของวิธี IF-LOF

จากรูปที่ ง.8 พบว่าข้อมูลส่วนใหญ่ของกลุ่มที่ 1 ข้อมูลมีการเกาะกลุ่มกันและมีการกระจายตัวที่น้อย และกลุ่มที่ 2 ข้อมูลมีการกระจายตัวที่มาก และค่า Silhouette Coefficient มีค่าเท่ากับ 0.8681 ซึ่งมีค่าใกล้ 1 แสดงว่าข้อมูลถูกจัดให้อยู่ในกลุ่มมีความเหมาะสมแล้ว

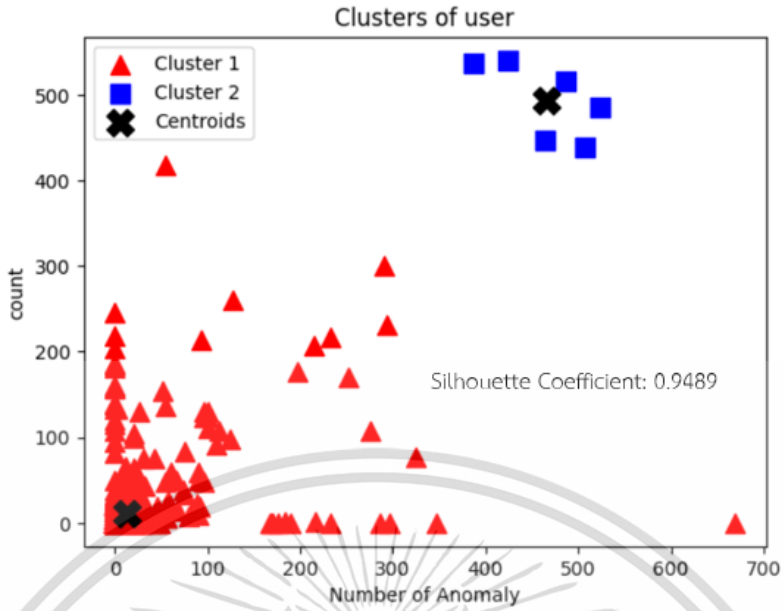
ง.1.2 ข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลว

ง.1.2.1 วิธีโลคอลเอาท์ไลเออร์แฟคเตอร์



รูปที่ ง.9 Elbow หาค่า K ที่เหมาะสม และค่า Silhouette ของวิธี LOF

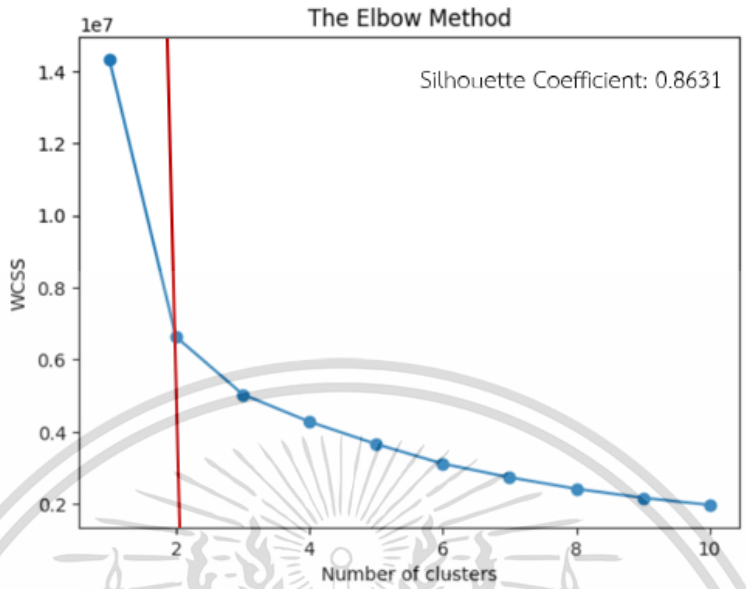
รูปที่ ง.9 จะพบว่าจำนวนกลุ่มที่เหมาะสมในการจัดกลุ่มคือ 2 กลุ่ม (K = 2) และค่า Silhouette Coefficient มีค่าเท่ากับ 0.9489 ซึ่งมีค่าใกล้ 1 แสดงว่าข้อมูลถูกจัดให้อยู่ในกลุ่มมีความเหมาะสมแล้ว ซึ่งถ้ามองจากภาพแผนภาพ Elbow ที่มีจุดหักศอกที่ 2 ก็มีความเป็นไปได้ที่ข้อมูลจะแบ่งเป็น 2 กลุ่มเช่นเดียวกัน แสดงว่าการจัดกลุ่ม 2 กลุ่มเหมาะสมกับข้อมูลชุดนี้ที่สุด จึงทำการจัดกลุ่มข้อมูล 2 กลุ่ม



รูปที่ ง.10 รูปการแบ่งกลุ่มของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลวของวิธี LOF

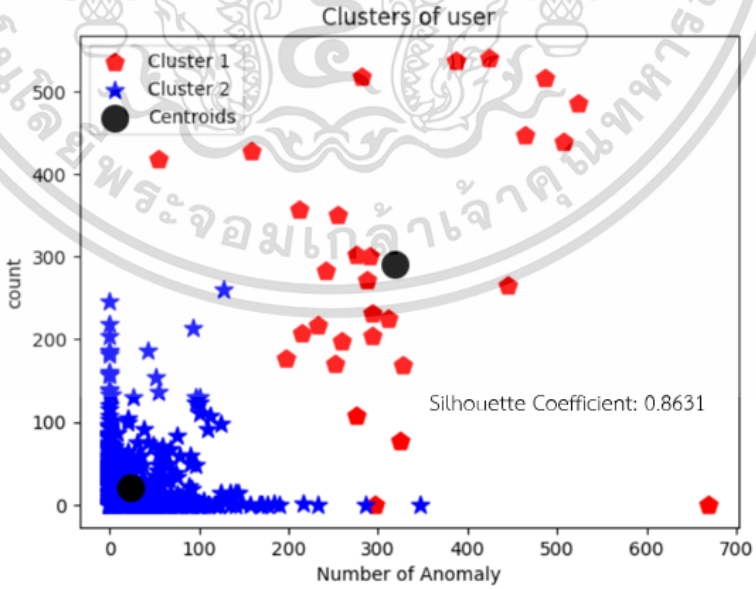
จากรูปที่ ง.10 พบว่าข้อมูลส่วนใหญ่ของกลุ่มที่ 1 และกลุ่มที่ 2 ข้อมูลมีการเกาะกลุ่มกันและมีการกระจายตัวที่น้อย และค่า Silhouette Coefficient มีค่าเท่ากับ 0.9489 ซึ่งมีค่าใกล้ 1 แสดงว่าข้อมูลถูกจัดให้อยู่ในกลุ่มมีความเหมาะสมแล้ว

ง.1.2.2 วิธีซัพพอร์ตเวกเตอร์แมชชีนคลาสหนึ่ง



รูปที่ ง.11 Elbow หาค่า K ที่เหมาะสม และค่า Silhouette ของวิธี OCSVM

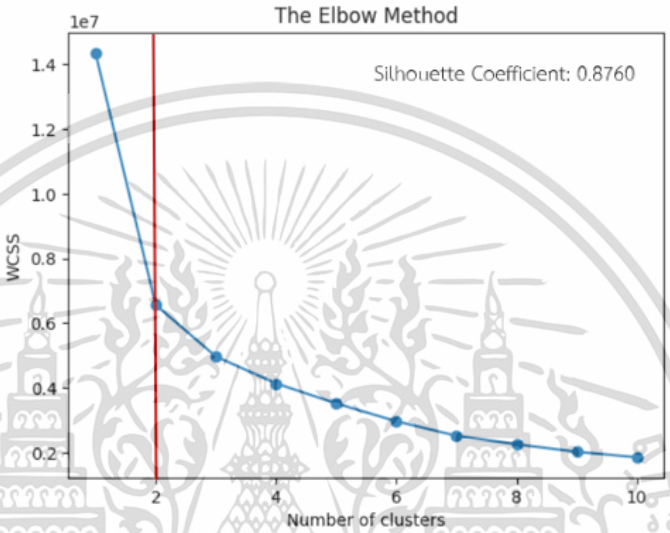
รูปที่ ง.11 จะพบว่าจำนวนกลุ่มที่เหมาะสมในการจัดกลุ่มคือ 2 กลุ่ม (K = 2) และค่า Silhouette Coefficient มีค่าเท่ากับ 0.8631 ซึ่งมีค่าใกล้ 1 แสดงว่าข้อมูลถูกจัดให้อยู่ในกลุ่มมีความเหมาะสมแล้ว ซึ่งถ้ามองจากภาพแผนภาพ Elbow ที่มีจุดหักศอกที่ 2 ก็มีความเป็นไปได้ที่ข้อมูลจะแบ่งเป็น 2 กลุ่มเช่นเดียวกัน แสดงว่าการจัดกลุ่ม 2 กลุ่มเหมาะสมกับข้อมูลชุดนี้ที่สุด จึงทำการจัดกลุ่มข้อมูล 2 กลุ่ม



รูปที่ ง.12 รูปการแบ่งกลุ่มของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลว ของวิธี OCSVM

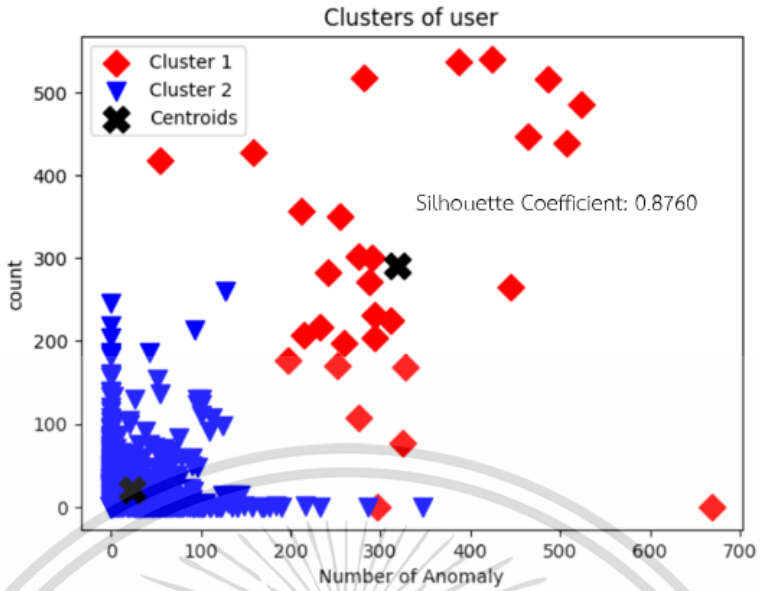
จากรูปที่ ง.12 พบว่าข้อมูลส่วนใหญ่ของกลุ่มที่ 1 ข้อมูลมีการเกาะกลุ่มกันและมีการกระจายตัวที่น้อยซึ่งแสดง และกลุ่มที่ 2 ข้อมูลมีการกระจายตัวกันเล็กน้อย และค่า Silhouette Coefficient มีค่าเท่ากับ 0.8631 ซึ่งมีค่าใกล้ 1 แสดงว่าข้อมูลถูกจัดให้อยู่ในกลุ่มมีความเหมาะสมแล้ว

ง.1.2.3 วิธีไอโซเลชั่นฟอเรส



รูปที่ ง.13 Elbow หาค่า K ที่เหมาะสม และค่า Silhouette ของวิธี OCSVM

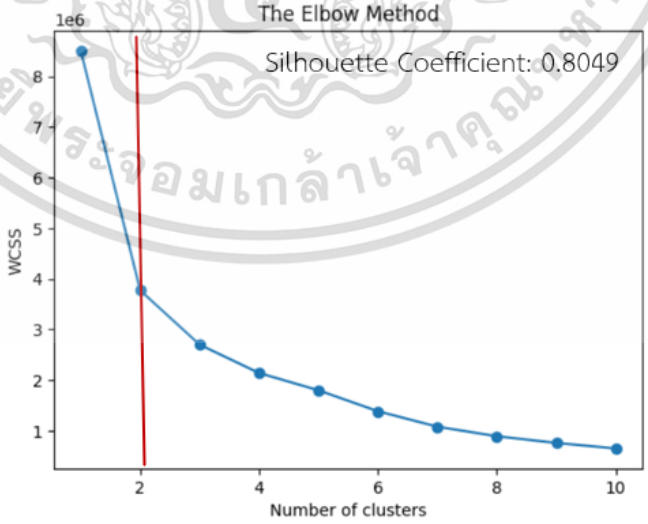
รูปที่ ง.13 จะพบว่าจำนวนกลุ่มที่เหมาะสมในการจัดกลุ่มคือ 2 กลุ่ม (K = 2) และค่า Silhouette Coefficient มีค่าเท่ากับ 0.8760 ซึ่งมีค่าใกล้ 1 แสดงว่าข้อมูลถูกจัดให้อยู่ในกลุ่มมีความเหมาะสมแล้ว ซึ่งถ้ามองจากภาพแผนภาพ Elbow ที่มีจุดหักศอกที่ 2 ก็มีความเป็นไปได้ที่ข้อมูลจะแบ่งเป็น 2 กลุ่มเช่นเดียวกัน แสดงว่าการจัดกลุ่ม 2 กลุ่มเหมาะสมกับข้อมูลชุดนี้ที่สุด จึงทำการจัดกลุ่มข้อมูล 2 กลุ่ม



รูปที่ ง.14 รูปการแบ่งกลุ่มของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลวของวิธี Isolation Forest

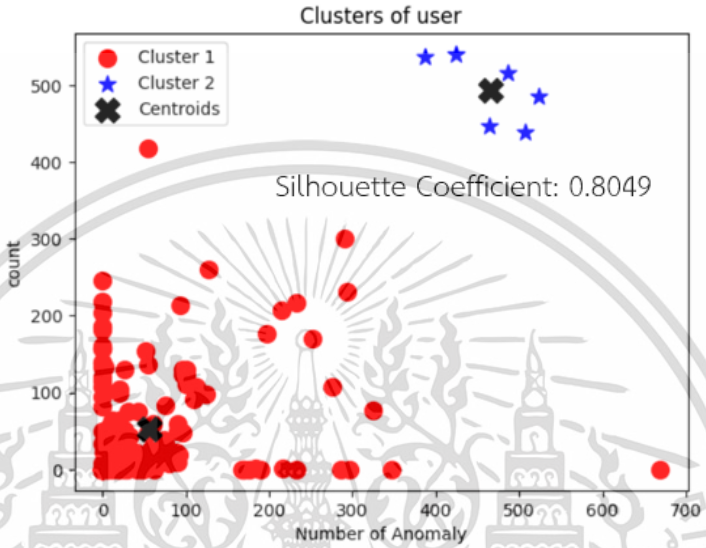
จากรูปที่ ง.14 พบว่าข้อมูลส่วนใหญ่ของกลุ่มที่ 1 ข้อมูลมีการเกาะกลุ่มกันและมีการกระจายตัวที่น้อยซึ่งแสดง และกลุ่มที่ 2 ข้อมูลมีการกระจายตัวกันเล็กน้อย และค่า Silhouette Coefficient มีค่าเท่ากับ 0.8760 ซึ่งมีค่าใกล้ 1 แสดงว่าข้อมูลถูกจัดให้อยู่ในกลุ่มมีความเหมาะสมแล้ว

ง.1.2.4 วิธีไอเอฟ-แอลโอเอฟ



รูปที่ ง.15 Elbow หาค่า K ที่เหมาะสม และค่า Silhouette ของวิธี IF-LOF

รูปที่ ง.15 จะพบว่าจำนวนกลุ่มที่เหมาะสมในการจัดกลุ่มคือ 2 กลุ่ม (K = 2) และค่า Silhouette Coefficient มีค่าเท่ากับ 0.8049 ซึ่งมีค่าใกล้ 1 แสดงว่าข้อมูลถูกจัดให้อยู่ในกลุ่มมีความเหมาะสมแล้ว ซึ่งถ้ามองจากภาพแผนภาพ Elbow ที่มีจุดหักศอกที่ 2 ก็มีความเป็นไปได้ที่ข้อมูลจะแบ่งเป็น 2 กลุ่มเช่นเดียวกัน แสดงว่าการจัดกลุ่ม 2 กลุ่มเหมาะสมกับข้อมูลชุดนี้ที่สุด จึงทำการจัดกลุ่มข้อมูล 2 กลุ่ม



รูปที่ ง.16 รูปการแบ่งกลุ่มของข้อมูลรายการเข้ามาใช้งานในระบบที่มีสถานะล้มเหลว ของวิธี IF-LOF

จากรูปที่ ง.16 พบว่าข้อมูลส่วนใหญ่ของกลุ่มที่ 1 และกลุ่มที่ 2 ข้อมูลมีการเกาะกลุ่มกันและมีการกระจายตัวที่น้อย และค่า Silhouette Coefficient มีค่าเท่ากับ 0.8049 ซึ่งมีค่าใกล้ 1 แสดงว่าข้อมูลถูกจัดให้อยู่ในกลุ่มมีความเหมาะสมแล้ว

ประวัติผู้วิจัย

พุฒิธร ชลิตชัยยะ (แม็ค)

ที่อยู่ : 8 ซ.จันทน์ 34 ถนน จันทน์ แขวงทุ่งวัดดอน เขตสาทร กรุงเทพ 10120
เบอร์โทรศัพท์ : 061-938-4191 Email : putthitorn.cha@gmail.com

ประวัติการศึกษา

- ระดับมัธยมศึกษา
  โรงเรียนพระแม่มารีสาทร จังหวัดกรุงเทพมหานคร
- ระดับปริญญาตรี
  วิทยาศาสตรบัณฑิต สาขาสถิติประยุกต์
  ภาควิชาสถิติ คณะวิทยาศาสตร์
  สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

งานทะเบียนคณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
คำรับรองเล่มสหกิจศึกษา
วันที่ 10 เดือน กรกฎาคม พ.ศ. 2566

ข้าพเจ้า นายพุฒิธร ชลิตชัยยะ รหัสนักศึกษา 62050806
นักศึกษาหลักสูตรวิทยาศาสตรบัณฑิต สาขาวิชา สถิติประยุกต์ ภาควิชา สถิติ ขอรับรองว่าโครงงานสหกิจศึกษา

เรื่อง ชื่อภาษาไทย	การตรวจจับความผิดปรกติของการเข้าใช้ระบบ Azure Active Directory ด้วยเทคนิคการเรียนรู้ของเครื่อง
ชื่อภาษาอังกฤษ	Detecting Logon Anomaly of Azure Active Directory Service using Machine Learning Techniques
ปีการศึกษา	2565

เป็นผลงานวิจัยที่มิได้คัดลอกหรือละเมิดลิขสิทธิ์ของผู้อื่นและได้ผ่านการตรวจสอบความซ้ำซ้อนเรียบร้อยแล้ว และได้ แนบเอกสารการตรวจสอบการลอกเลียนงานวรรณกรรมที่ตรวจสอบจากเล่มโครงงานสหกิจศึกษาฉบับสมบูรณ์แล้ว โปรแกรมอักขราวิสุทธิ์ 2.20 % โปรแกรม Turnitin 22 %

ลงชื่อ.....
(นายพุฒิธร ชลิตชัยยะ)
นักศึกษา

ข้าพเจ้า ผศ.ดร.พรพิมล ชัยวุฒิศักดิ์ อาจารย์ที่ปรึกษาโครงงานสหกิจศึกษา ได้ตรวจสอบโครงงานสหกิจศึกษาของ นักศึกษาข้างต้น แล้ว ขอรับรองว่าเป็นผลงานวิจัยของนักศึกษาจริงและมีเนื้อหาสมบูรณ์ จึงลงชื่อไว้เป็นหลักฐาน

ลงชื่อ.....
(ผศ.ดร.พรพิมล ชัยวุฒิศักดิ์)
อาจารย์ที่ปรึกษา