การปรับปรุงวิศวกรรมคุณลักษณะสำหรับตัวจำแนกประเภทแบบต้นไม้สำหรับการตรวจจับการบุกรุกทางเครือข่าย

กิตติภพ มหาวัน

วิทยานิพนธ์เสนอบัณฑิตวิทยาลัย มหาวิทยาลัยนเรศวร
เพื่อเป็นส่วนหนึ่งของการศึกษา หลักสูตรปรัชญาดุษฎีบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์
ปีการศึกษา 2565

การปรับปรุงวิศวกรรมคุณลักษณะสำหรับตัวจำแนกประเภทแบบต้นไม้สำหรับการตรวจจับการบุกรุกทางเครือข่าย

กิตติภพ มหาวัน

วิทยานิพนธ์เสนอบัณฑิตวิทยาลัย มหาวิทยาลัยนเรศวร
เพื่อเป็นส่วนหนึ่งของการศึกษา หลักสูตรปรัชญาดุษฎีบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์
ปีการศึกษา 2565

วิทยานิพนธ์ เรื่อง "การปรับปรุงวิศวกรรมคุณลักษณะสำหรับตัวจำแนกประเภทแบบต้นไม้สำหรับการตรวจจับการบุกรุกทางเครือข่าย"

ของ กิตติภพ มหาวัน

ได้รับการพิจารณาให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการสอบวิทยานิพนธ์

(ผู้ช่วยศาสตราจารย์ ดร.สาคร เมฆรักษาวนิช)

..... ประธานที่ปรึกษาวิทยานิพนธ์

(ผู้ช่วยศาสตราจารย์ ดร.วินัย วงษ์ไทย)

..... กรรมการผู้ทรงคุณวุฒิภายใน

(รองศาสตราจารย์ ดร.จักรกฤษณ์ เสน่ห์ นมะหุต)

..... กรรมการผู้ทรงคุณวุฒิภายใน

(รองศาสตราจารย์ ดร.ไกรศักดิ์ เกษร)

..... กรรมการผู้ทรงคุณวุฒิภายใน

(ผู้ช่วยศาสตราจารย์ ดร.ธนะธร พ่อค้า)

..... กรรมการผู้ทรงคุณวุฒิภายใน

(ผู้ช่วยศาสตราจารย์ ดร.จันทร์จิรา พยัคฆ์เพศ)

อนุมัติ

.....

(รองศาสตราจารย์ ดร.กรองกาญจน์ ชูทิพย์)

คณบดีบัณฑิตวิทยาลัย

ชื่อเรื่อง การปรับปรุงวิศวกรรมคุณลักษณะสำหรับตัวจำแนกประเภทแบบต้นไม้สำหรับการตรวจจับการบุกรุกทางเครือข่าย

ผู้วิจัย กิตติภพ มหาวัน

ประธานที่ปรึกษา ผู้ช่วยศาสตราจารย์ ดร.วินัย วงษ์ไทย

ประเภทสารนิพนธ์ วิทยานิพนธ์ ปร.ด. วิทยาการคอมพิวเตอร์, มหาวิทยาลัยนเรศวร, 2565

คำสำคัญ ระบบตรวจจับการบุกรุกเครือข่าย, ลักษณนามแบบต้นไม้, วิศวกรรมคุณลักษณะ, อัลกอริทึมทางพันธุกรรม, การเลือกคุณสมบัติ การสร้างคุณสมบัติพหุนาม

บทคัดย่อ

วิทยานิพนธ์ที่นำเสนอเสนอวิธีการปรับปรุงระบบตรวจจับการบุกรุก (IDS) โดยใช้การเลือกคุณลักษณะ การสร้างคุณลักษณะพหุนาม และอัลกอริธึมเชิงพันธุกรรม วัตถุประสงค์หลักคือเพื่อปรับปรุงความแม่นยำของ IDS ในการตรวจจับและบรรเทาภัยคุกคามทางไซเบอร์ วิธีการวิจัยเกี่ยวข้องกับการรวบรวมล็อกไฟล์และข้อมูลเครือข่ายจากระบบเป้าหมาย ตามด้วยการประมวลผลล่วงหน้าเพื่อสร้างแบบจำลองมาตรฐานในการระบุคุณสมบัติที่ให้ข้อมูลมากที่สุดสำหรับการตรวจจับการบุกรุก ใช้วิธีการเลือกคุณสมบัติหลายวิธี รวมถึงการวิเคราะห์ไคสแควร์ การวิเคราะห์ช่องโหว่ความแปรปรวน และข้อมูลร่วม เทคนิคเหล่านี้ช่วยในการกำหนดลักษณะที่มีส่วนช่วยในการตรวจจับการบุกรุกได้มากที่สุด นอกจากนี้ การสร้างคุณสมบัติพหุนามยังใช้เพื่อบันทึกความสัมพันธ์และการโต้ตอบที่ไม่ใช่เชิงเส้นระหว่างคุณสมบัติต่างๆ ภายในอัลกอริทึม อัลกอริทึมทางพันธุกรรมถูกนำมาใช้เพื่อเพิ่มประสิทธิภาพทั้งกระบวนการเลือกคุณสมบัติและกระบวนการสร้างคุณสมบัติพหุนาม ด้วยการใช้จีโนไทป์ อัลกอริทึมพยายามระบุคุณสมบัติที่เกี่ยวข้องมากที่สุดและสร้างคุณสมบัติพหุนามที่ปรับปรุงความแม่นยำของ IDS แนวทางที่เสนอได้รับการประเมินโดยใช้ชุดข้อมูลที่เป็นมาตรฐานและเปิดเผยต่อสาธารณะ ซึ่งแสดงให้เห็นถึงการปรับปรุงที่สำคัญในความถูกต้องของ IDS ในขณะที่ลดข้อกำหนดในการจัดเก็บข้อมูลสำหรับกิจกรรมเครือข่ายได้อย่างมีประสิทธิภาพ ผลการวิจัยนี้มีนัยสำคัญสำหรับความปลอดภัยทางไซเบอร์ วิธีแก้ปัญหาเชิงปฏิบัติที่เสนอในวิทยานิพนธ์นี้มีส่วนช่วยในการเพิ่มประสิทธิภาพการทำงานของ IDS และแก้ไขข้อจำกัดของพื้นที่เก็บข้อมูลเพื่อปรับปรุงความสามารถของ IDS ข้อมูลเชิงลึกที่ได้รับจากการวิจัยนี้สามารถแจ้งการออกแบบระบบตรวจจับการบุกรุกที่มีประสิทธิภาพและประสิทธิผลมากขึ้น ในขณะที่ปรับการใช้พื้นที่เก็บข้อมูลให้เหมาะสม

สำหรับกิจกรรมเครือข่าย

Title THE MODIFIED FEATURE ENGINEERING APPROACH FOR TREE-BASED CLASSIFIERS OF NETWORK INTRUSION DETECTION

Author Kittiphop Mahawan

Advisor Assistant Professor Winai Wongthai, Ph.D.

Academic Paper Ph.D. Dissertation in Computer Science - (Type 2.1), Naresuan University, 2022

Keywords Log File, Hyperparameter tuning, Framework, Feature Selection, Polynomial Feature generation, Tree Based Classifier

ABSTRACT

The presented thesis proposes a method to enhance intrusion detection systems (IDS) through the use of feature selection, polynomial feature generation, and genetic algorithms. The primary objective is to improve the accuracy of IDS in detecting and mitigating cyber threats. The research methodology involves the collection of log files and network data from the target system, followed by preprocessing to establish a standard model.To identify the most informative features for intrusion detection, several feature selection methods were employed, including chi-squared analysis, vulnerability analysis, variance, and joint information. These techniques aid in determining the characteristics that contribute the most to the detection of intrusions. Moreover, polynomial feature generation was utilized to capture non-linear relationships and interactions among features within the algorithm.Genetic algorithms were employed to optimize both the feature selection and polynomial feature generation processes. By using genotypes, the algorithm sought to identify the most relevant features and generate polynomial properties that enhance the accuracy of IDS. The proposed approach was evaluated using standardized and publicly available datasets, demonstrating a significant improvement in IDS accuracy while effectively reducing storage requirements for network activity. The findings of this research have valuable implications for the field

of cybersecurity. The practical solutions proposed in this thesis contribute to enhancing the performance of IDS and addressing storage constraints in order to improve IDS capabilities. The insights provided by this research can inform the design of more efficient and effective intrusion detection systems, while optimizing storage usage for network activity.

ประกาศคุณูปการ

ผู้วิจัยขอกราบขอบพระคุณเป็นอย่างสูงในความกรุณาของ ผู้ช่วยศาสตราจารย์ ดร.วินัย วงษ์ไทย ประธานที่ปรึกษาวิทยานิพนธ์ที่ได้อุตส่าห์สละเวลาอันมีค่ามาเป็นที่ปรึกษา พร้อมทั้งให้คำแนะนำตลอดเวลาในการทำวิทยานิพนธ์ฉบับนี้ และกราบขอบพระคุณคณะกรรมการสอบวิทยานิพนธ์อันประกอบไปด้วย ผู้ช่วยศาสตราจารย์ ดร.สาคร เมฆรักษาวนิช ประธานกรรมการสอบ ป้องกันวิทยานิพนธ์ รองศาสตราจารย์ ดร.จักรกฤษณ์ เสน่ห์ นมะหุต รองศาสตราจารย์ ดร.ไกรศักดิ์ เกษร ผู้ช่วยศาสตราจารย์ ดร.ธนะธร พ่อค้า และ ผู้ช่วยศาสตราจารย์ ดร.จันทร์จิรา พยัคฆ์เพศ กรรมการผู้ทรงคุณวุฒิภายใน ที่ได้กรุณาให้คำแนะนำตลอดจนแก้ไขข้อบกพร่องของวิทยานิพนธ์ด้วยความเอาใจใส่ จนทำให้วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้อย่างสมบูรณ์และทรงคุณค่าเหนือสิ่งอื่นใดขอกราบขอบพระคุณ บิดา มารดาและเพื่อนของผู้วิจัยที่ให้กำลังใจและให้ การสนับสนุนในทุกๆ ด้านอย่างดีที่สุดเสมอมาคุณค่าและคุณประโยชน์อันพึงจะมีจากวิทยานิพนธ์ฉบับนี้ ผู้วิจัยขอมอบและอุทิศแด่ผู้มีพระคุณทุกๆ ท่าน ผู้วิจัยหวังเป็นอย่างยิ่งว่า งานวิจัยนี้จะเป็นประโยชน์ต่อผู้ที่สนใจนำไปใช้ประโยชน์ไม่มากก็น้อย หากมีข้อบกพร่องประการใดที่อาจจะเกิดขึ้นภายในวิทยานิพนธ์ ผู้วิจัยขอน้อมรับเพื่อเป็นประโยชน์ในการพัฒนางานวิจัยต่อไป

กิตติภพ มหาวัน

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

บทที่ 1

บทนำ

เนื้อหาในบทที่ 1 ผู้วิจัยขอกล่างถึงในภาพรวมของเรื่องราวทั้งหมดในวิทยานิพนธ์ เรื่อง การปรับปรุงวิศวกรรมคุณลักษณะสำหรับตัวจำแนกประเภทแบบต้นไม้สำหรับการตรวจจับการบุกรุกทางเครือข่าย เพื่อให้ผู้อ่านจะได้เข้าใจง่ายก่อนที่จะได้ทำการศึกษาในรายละเอียดเชิงลึกต่อไป โดยในบทที่ 1 จะประกอบไปด้วยเนื้อหาและรายละเอียด โดยแบ่งแยกตามหัวข้อดังนี้

1. ความเป็นมาและความสำคัญของปัญหา
2. ปัญหาวิจัย
3. เป้าหมายของงานวิจัย
4. วัตถุประสงค์ของงานวิจัย
5. ขอบเขตของงานวิจัย
6. ประโยชน์ที่คาดว่าจะได้รับ

ความเป็นมาและความสำคัญของปัญหา

การใช้งานระบบคอมพิวเตอร์ผ่านระบบเครือข่ายนั้นจะมีการจัดเก็บข้อมูลของคอมพิวเตอร์ที่เชื่อมต่อกับเครือข่าย โดยจะถูกจัดเก็บไว้ที่เครื่องของผู้ให้บริการ ซึ่งเรียกว่า ข้อมูลจราจรทางคอมพิวเตอร์ (Log File) โดยในงานวิทยานิพนธ์นี้เรียกว่า ล็อกไฟล์ ในการบันทึกของล็อกไฟล์นั้น จำแนกได้เป็นหลายประเภท (Abdalla & Jumaa, 2022) เช่น การบันทึกแอปพลิเคชัน (Application Log), การบันทึกเว็บเซิร์ฟเวอร์ (Web Server Log), การบันทึกระบบ (System Log), การบันทึกความปลอดภัย (Security Log), การบันทึกเครือข่าย (Network Log) เป็นต้น การบันทึกความปลอดภัยที่นำมาซึ่งการรวบรวมข้อมูลจากระบบที่เกี่ยวข้องกับความปลอดภัย และช่วยในการระบุการละเมิด โปรแกรมที่เป็นอันตราย การโจรกรรมข้อมูล และเพื่อประเมินสภาพของมาตรการรักษาความปลอดภัย ล็อกไฟล์การเข้าถึงนั้นจะมีข้อมูลเกี่ยวกับการรับรองความถูกต้องของผู้ใช้ ความล้มเหลวของระบบและความผิดปกติจะถูกบันทึกอยู่ในล็อกไฟล์เหล่านี้ด้วย (David et al., 2022) จากพระราชบัญญัติว่าด้วยการกระทำความผิดเกี่ยวกับคอมพิวเตอร์ (ฉบับที่ 2) พ.ศ. 2560 “มาตรา 26 ผู้ให้บริการต้องเก็บรักษาข้อมูลจราจรทางคอมพิวเตอร์ไว้ไม่น้อยกว่าเก้าสิบวัน นับแต่วันที่ข้อมูลนั้นเข้าสู่ระบบคอมพิวเตอร์แต่ในกรณีจำเป็น พนักงานเจ้าหน้าที่จะสั่งให้ผู้ให้บริการผู้ใดเก็บรักษาข้อมูลจราจรทางคอมพิวเตอร์ไว้เกินเก้าสิบวันแต่ไม่เกินสองปีเป็นกรณีพิเศษเฉพาะรายและเฉพาะคราวก็ได้” ซึ่งข้อมูลล็อกไฟล์เหล่านี้จะสามารถนำไปใช้ในการตรวจสอบหาผู้กระทำความผิดได้ โดย

ข้อมูลล็อกไฟล์นี้มีปริมาณข้อมูลที่สูงมากเมื่อมีผู้ใช้งานหลายคนจะทำให้ปริมาณข้อมูลสูงขึ้นไปด้วยซึ่งเป็นการสิ้นเปลืองพื้นที่ในการจัดเก็บข้อมูลล็อกไฟล์ในเครื่องคอมพิวเตอร์ สำหรับข้อมูลล็อกไฟล์นั้นสามารถนำไปสู่การตรวจจับการบุกรุกทางเครือข่ายได้ด้วยการนำล็อกไฟล์มาใช้ในการวิเคราะห์เพื่อหาความผิดปกตินั้นได้มีนักวิจัยได้ทำการวิจัยเกี่ยวกับล็อกไฟล์ เช่น Ertam & Kaya (2018), Brandao & Georgieva (2020), Ryciak et al. (2022) และ Wadekar et al. (2019) ได้มีการนำข้อมูลจากล็อกไฟล์มาใช้ในการวิเคราะห์เพื่อหาความผิดปกติของการทำงานของระบบคอมพิวเตอร์ที่อาจจะถูกผู้ไม่ประสงค์ดีเข้ามาในระบบการทำงานด้วยวิธีการการเรียนรู้ของเครื่อง (Machine Learning) ด้วยเทคนิคการสร้างตัวแบบต่าง ๆ เพื่อหาประสิทธิภาพของตัวแบบที่เหมาะสมสำหรับข้อมูลนั้น ๆ การที่จะวิเคราะห์ความผิดปกติในการสร้างตัวแบบนั้นการคัดเลือกคุณลักษณะของข้อมูลและการกำหนดค่าไฮเปอร์พารามิเตอร์ของแต่ละตัวแบบเป็นสิ่งที่สามารถช่วยเพิ่มประสิทธิภาพการทำงานของตัวแบบให้มีความถูกต้องได้ ซึ่งการหาคุณลักษณะที่สำคัญนั้นไม่เพียงแต่ช่วยในเรื่องความเร็วของการจัดการข้อมูลเท่านั้นแต่ยังช่วยในเรื่องของการปรับปรุงอัตราการตรวจจับได้ (Brandao & Georgieva, 2020) จึงส่งผลให้การคัดเลือกคุณลักษณะนั้นมีความสำคัญเพิ่มมากขึ้นสำหรับการวิเคราะห์ข้อมูล (Data Analysis) การเรียนรู้ของเครื่อง และการทำเหมืองข้อมูล (Data Mining) โดยเฉพาะอย่างชุดข้อมูลที่มีมิติของข้อมูลสูงจึงจำเป็นจะต้องมีการคัดกรองคุณลักษณะที่ไม่เกี่ยวข้องกันและซ้ำซ้อนกัน (Bommert et al., 2020) เพื่อลดผลกระทบด้านมิติข้อมูลในชุดข้อมูลผ่านการค้นหาชุดย่อยของคุณสมบัติที่กำหนดของข้อมูลได้อย่างมีประสิทธิภาพ (Zebari et al., 2020) โดยจุดประสงค์หลักของการคัดเลือกคุณลักษณะคือการสร้างส่วนย่อยของคุณลักษณะให้มีขนาดเล็กที่สุดเท่าที่จะเป็นไปได้ แต่ยังคงแสดงถึงคุณลักษณะที่สำคัญของข้อมูลอินพุตทั้งหมด (Velliangiri& Alagumuthukrishnan, 2019, Eesa et al., 2015) ซึ่งการคัดเลือกคุณลักษณะจะช่วยในการลดขนาดของข้อมูล ลดพื้นที่ในการจัดเก็บข้อมูล ปรับปรุงความถูกต้องในการพยากรณ์ การหลีกเลี่ยงพฤติกรรมการเรียนรู้ของเครื่องที่ไม่พึงปรารถนา (Overfitting) การลดเวลาในการประมวลผล โดยเมื่อนำวิธีการคัดเลือกคุณลักษณะไปใช้ในการคัดเลือกคุณลักษณะของล็อกไฟล์จะทำให้ปริมาณการเก็บข้อมูลของล็อกไฟล์มีขนาดเล็กลง และยังคงเก็บคุณลักษณะที่สำคัญที่สามารถบ่งบอกถึงความผิดปกติจากการทำงานได้โดยที่ความถูกต้องการวิเคราะห์เพื่อหาความผิดปกติของการทำงานของระบบคอมพิวเตอร์ไม่เปลี่ยนแปลงไปจากเดิม ในวิธีการทำการคัดเลือกคุณลักษณะสามารถดำเนินการได้หลายรูปแบบโดยเฉพาะอย่างยิ่งวิธีการฟิลเตอร์ (Filter) โดยใช้การคำนวณค่าทางสถิติมาให้ในการประเมินระดับความสำคัญของคุณลักษณะย่อย (Subset Feature) ซึ่งเทคนิคนี้ให้ประสิทธิภาพที่ดีและการประมวลผลที่มีประสิทธิภาพสูง ปรับขนาดได้ง่ายสำหรับชุดข้อมูลที่มีมิติสูง (Zebari et al., 2020)

การคัดเลือกคุณลักษณะสามารถใช้การสร้างคุณลักษณะพหุนาม (Polynomial Feature Generation) ซึ่งเทคนิคที่ใช้เพื่อสร้างคุณลักษณะใหม่โดยการดำเนินการทางคณิตศาสตร์กับคุณลักษณะที่มีอยู่ด้วยการเพิ่มเป็นเลขชี้กำลัง (Ostertagová, 2012) ในการหาคุณลักษณะที่สำคัญ วิธีการการสร้างคุณลักษณะพหุนามจะเป็นการหาคุณลักษณะที่ซ่อนอยู่ในชุดข้อมูลเพื่อหาความสำคัญที่ซ่อนอยู่ของคุณลักษณะนั้นเพื่อนำมาใช้ประโยชน์ในการวิเคราะห์เพื่อหาความผิดปกติของการทำงานของระบบคอมพิวเตอร์ เช่น งานวิจัยของ Sciavicco et al. (2021) ได้ทำการตรวจสอบวิธีการแบบจำลองจากล่างขึ้นบน (Bottom-Up Approach) โดยใช้อัลกอริธึม Polynomial Time ในการดึงข้อมูล Conditional Simple Temporal Network with Uncertainty and Decisions (CSTNUD) จากชุดของการติดตามการดำเนินงาน เช่น Log งานวิจัยของ Vinaroz et al. (2022) ได้เสนอวิธีการที่จะแทนที่คุณสมบัติแบบสุ่มด้วยคุณสมบัติแบบ Hermite Polynomial ซึ่งสามารถประมาณค่าเฉลี่ยการฝังการกระจายข้อมูลได้อย่างแม่นยำเมื่อเทียบกับคุณสมบัติแบบสุ่ม

ขั้นตอนวิธีเชิงพันธุกรรม เป็นอีกหนึ่งวิธีการที่ใช้ในการค้นหาคุณลักษณะของชุดข้อมูล และการหาค่าไฮเปอร์พารามิเตอร์ที่เหมาะสม โดยสามารถค้นหาชุดของวิธีการที่หลากหลายด้วยเทคนิคการค้นหาตามหลักการวิวัฒนาการของการคัดเลือกโดยธรรมชาติและพันธุศาสตร์ เนื่องจากความสามารถในการค้นหาบริเวณต่าง ๆ ในพื้นที่ของวิธีการที่ต้องการค้นหา (Khotimah et al., 2020) Gharaee & Hosseinvand. (2016) ใช้วิธีการคัดเลือกคุณลักษณะด้วยพื้นฐานของพันธุกรรมด้วยนวัตกรรมฟิตเนสฟังก์ชันที่ลดมิติของข้อมูลโดยเพิ่มผลการตรวจจับที่ตรงกับสิ่งที่ต้องการตรวจจับของระบบการตรวจจับการบุกรุก (Intrusion Detection System หรือ IDS) และได้ใช้ขั้นตอนวิธีเชิงพันธุกรรมร่วมกับซัพพอร์ตเวคเตอร์แมชชีน (Support Vector Machine หรือ SVM) ในการตรวจจับความผิดปกติ Zhao et al. (2011) ใช้ขั้นตอนวิธีเชิงพันธุกรรมเพื่อปรับค่าพารามิเตอร์ของ SVM ให้เหมาะสม และได้ใช้ขั้นตอนวิธีเชิงพันธุกรรมและโครโมโซมของคุณลักษณะและพารามิเตอร์ SVM เพื่อเพิ่มประสิทธิภาพให้กับคุณลักษณะและพารามิเตอร์ ด้วยการทดลองกับชุดข้อมูลการตรวจหาโรคหัวใจและโรคมะเร็งจากชุดข้อมูลของ UCI โดยฟิตเนสฟังก์ชันช่วยให้โครโมโซมมีความแม่นยำในการจำแนกสูงสุดและจำนวนคุณลักษณะน้อยที่สุด

จากความเป็นมาและความสำคัญของปัญหาในข้างต้นการคัดเลือกคุณลักษณะและการหาค่าไฮเปอร์พารามิเตอร์นั้นมีการใช้วิธีการด้วยการสร้างคุณลักษณะพหุนาม หรือขั้นตอนวิธีเชิงพันธุกรรม ซึ่งยังไม่มีงานวิจัยที่นำสองวิธีการนี้มาทำงานร่วมกัน ในงานวิทยานิพนธ์นี้จึงมุ่งเน้นในการพัฒนาเฟรมเวิร์คสำหรับการค้นหาการคัดเลือกคุณลักษณะของข้อมูลและค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมที่สุดของชุดข้อมูลล็อกไฟล์ด้วยการทำงานของวิธีการการสร้างคุณลักษณะพหุนามนำมาร่วมกับขั้นตอนวิธีเชิงพันธุกรรมเพื่อลดขนาดและมิติของข้อมูลในการจัดเก็บล็อกไฟล์ตามพระราชบัญญัติว่าด้วยการ

กระทำความผิดเกี่ยวกับคอมพิวเตอร์ (ฉบับที่ 2) พ.ศ. 2560 ที่ระบุว่าผู้ให้บริการอินเทอร์เน็ตต้องจัดเก็บข้อมูลล็อกไฟล์ไว้ไม่ต่ำกว่า 90 วัน ซึ่งหลังจาก 90 วันข้อมูล ล็อกไฟล์นั้นจะยังคงอยู่ด้วยจำนวนคุณลักษณะของข้อมูลที่ลดลง นำเสนอวิธีการค้นหาและแสดงผลลัพธ์ที่หลากหลายเพื่อให้ผู้ใช้งานสามารถนำไปประยุกต์ใช้งานได้ และนำเสนอแนวทางการวิเคราะห์การหาจำนวนคุณลักษณะที่เหมาะสม และนำคุณลักษณะที่ได้มาทำการสร้างตัวแบบด้วยขั้นตอนวิธีการจำแนกประเภทแบบแรนดอมฟอร์เรสต์ (Random Forest) ในการทดสอบหาค่าความถูกต้อง (Accuracy) โดยเฟรมเวิร์คนี้จะมีความยืดหยุ่นที่สามารถปรับเปลี่ยนการทำงานของแต่ละขั้นตอนได้ตามลักษณะของข้อมูล

ปัญหาวิจัย

จากทั้งหมดที่กล่าวมาในหัวข้อความเป็นมาและความสำคัญของปัญหานั้น สามารถสรุปปัญหาในการวิจัยได้ดังนี้

1. จากงานวิจัยของ Ertam & Kaya (2018), Brandao & Georgieva (2020), Ryciak et al. (2022) และ Wadekar et al. (2019) ได้มีการนำข้อมูลจากล็อกไฟล์มาใช้ในการวิเคราะห์เพื่อหาความผิดปกติของการทำงานของระบบคอมพิวเตอร์ที่อาจจะถูกผู้ไม่ประสงค์ดีเข้ามาในระบบการทำงานด้วยวิธีการการเรียนรู้ของเครื่องด้วยเทคนิคการสร้างตัวแบบต่าง ๆ เพื่อหาประสิทธิภาพของตัวแบบที่เหมาะสมสำหรับข้อมูลนั้น ๆ โดยได้ทำการคัดเลือกคุณลักษณะที่มีความเหมาะสมที่จะให้ค่าการพยากรณ์ของตัวแบบมีความถูกต้องมากที่สุด ประกอบกับงานวิจัยของ Saha et al. (2021), Zhang et al. (2019) และ Li et al. (2022) ได้นำวิธีการพหุนามมาใช้ในการเพิ่มประสิทธิภาพความแม่นยำของตัวแบบด้วยการขยายคุณลักษณะที่มีความสำคัญซ่อนอยู่ แต่ยังไม่มีการพัฒนาเฟรมเวิร์คในการคัดเลือกคุณลักษณะและการสร้างคุณลักษณะพหุนาม

2. จากปัญหาการวิจัยในข้อ 1 และงานวิจัย Hosseinvand. (2016), Zhao et al. (2011) ยังไม่มีการนำวิธีการสร้างคุณลักษณะพหุนามมาทำงานร่วมกับขั้นตอนวิธีเชิงพันธุกรรมสำหรับการคัดเลือกคุณลักษณะและค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมที่สุดของชุดข้อมูลล็อกไฟล์

เป้าหมายของงานวิจัย

พัฒนาเฟรมเวิร์คสำหรับการค้นหาการคัดเลือกคุณลักษณะของข้อมูลและค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมที่สุดของชุดข้อมูลล็อกไฟล์ด้วยการทำงานของวิธีการการสร้างคุณลักษณะพหุนามนำมาร่วมกับขั้นตอนวิธีเชิงพันธุกรรมเพื่อลดขนาดและมิติของข้อมูลในการจัดเก็บล็อกไฟล์ นำเสนอวิธีการค้นหาและแสดงผลลัพธ์ที่หลากหลายเพื่อให้ผู้ใช้งานสามารถนำไปประยุกต์ใช้งานได้ และนำเสนอแนวทางการวิเคราะห์การหาจำนวนคุณลักษณะที่เหมาะสม

วัตถุประสงค์ของงานวิจัย

1. เพื่อพัฒนาเฟรมเวิร์คในการเลือกแบบจำลองประเภทแรนดอมฟอร์เรสต์ สำหรับปัญหาการจำแนกประเภทด้วยวิธีการคัดเลือกคุณลักษณะและการสร้างคุณลักษณะพหุนาม

2. เพื่อพัฒนาขั้นตอนวิธีสำหรับการคัดเลือกคุณลักษณะและการสร้างคุณลักษณะพหุนามด้วยขั้นตอนวิธีเชิงพันธุกรรม

3. เพื่อประยุกต์ใช้เฟรมเวิร์คหรือขั้นตอนวิธีสำหรับวัตถุประสงค์ในการทดสอบการลดพื้นที่การจัดเก็บเหตุการณ์ที่เกิดขึ้นในระบบเครือข่าย

ขอบเขตของงานวิจัย

1. ขอบเขตด้านกระบวนการ

1.1 ใช้แบบจำลอง Random Forest Classification เป็นตัวแทนของอัลกอริทึมตระกูล Tree base สำหรับการคัดเลือกคุณลักษณะและการสร้างคุณสมบัติพหุนาม

1.2 การคัดเลือกคุณลักษณะ ใช้ ฟิลเตอร์เมธอด และ แรปเปอร์เมธอด

1.3 การสร้างคุณสมบัติพหุนาม ใช้ดีกรีสูงสุดที่ 3 ดีกรี

1.4 ชุดข้อมูลในการทดลอง ใช้ 4 ชุดข้อมูล ประกอบด้วย 2 ชุดข้อมูลที่ไม่เกี่ยวข้องกับเครือข่ายและ 2 ชุดข้อมูลที่เกี่ยวข้องกับเครือข่าย

1.5 ใช้ความแม่นยำเป็นตัวแทนการวัดประสิทธิภาพของแบบจำลอง

1.6 ใช้แบบจำลองแรนดอมฟอร์เรสที่ไม่ผ่านการคัดเลือกคุณสมบัติและการสร้างพหุนามเป็น Base line

2. ขอบเขตด้านเทคโนโลยีและอุปกรณ์ซอฟต์แวร์

2.1 เครื่องคอมพิวเตอร์ CPU 12th Gen Intel(R) Core(TM) i9-12900KS 3.42 GHz, SDRAM 64 GB และ HDD 1 TB 1 เครื่อง ติดตั้งระบบปฏิบัติการ Windows10 ขนาด 64 bits สำหรับการทดลอง

2.2 ติดตั้งซอฟต์แวร์ Microsoft Visual Studio สำหรับการเขียนโปรแกรมภาษา Python สำหรับทำการทดสอบการทำงานของตัวแบบในการทดลอง

คุณูปการของงานวิจัย

1.ได้เฟรมเวิร์คสำหรับการคัดเลือกคุณสมบัติที่เหมาะสมและแบบจำลองที่เป็นชนิดต้นไม้ที่มีความแม่นยำสูง โดยไม่ยึดติดกับชุดข้อมูลใดชุดข้อมูลหนึ่งด้วยเทคนิคการสร้างคุณสมบัติแบบพหุนาม

2.ได้ขั้นตอนวิธีสำหรับการคัดเลือกคุณสมบัติที่เหมาะสมและแบบจำลองที่เป็นชนิดต้นไม้ที่มีความแม่นยำสูง โดยไม่ยึดติดกับจำนวนคุณสมบัติหรือชุดข้อมูลใดชุดข้อมูลหนึ่งด้วยเทคนิคการสร้างคุณสมบัติแบบพหุนามร่วมกับขั้นตอนวิธีเชิงพันธุกรรม

ประโยชน์ที่ได้รับ

1. ได้เฟรมเวิร์คในการคัดเลือกคุณสมบัติและค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมกับชุดข้อมูลในการตรวจจับความผิดปกติในล็อกไฟล์

2. เพื่อลดขนาดและมิติของข้อมูลในการจัดเก็บล็อกไฟล์ที่ส่งผลต่อการลดการใช้ทรัพยากรในการจัดเก็บข้อมูลล็อกไฟล์ของผู้ให้บริการ

3. นำเสนอวิธีการค้นหาและแสดงผลลัพธ์ที่หลากหลายเพื่อให้ผู้ใช้งานสามารถนำไปประยุกต์ใช้งานได้ และนำเสนอแนวทางการวิเคราะห์การหาจำนวนคุณลักษณะที่เหมาะสม

บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

ในงานวิจัย การปรับปรุงวิศวกรรมคุณลักษณะสำหรับตัวจำแนกประเภทแบบต้นไม้สำหรับการตรวจจับการบุกรุกทางเครือข่าย ผู้วิจัยได้ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องโดยจะมีเทคนิคและงานวิจัยที่เกี่ยวข้องตามหัวข้อต่อไปนี้

1. ล็อกไฟล์
2. การคัดเลือกคุณลักษณะ
3. การสร้างคุณลักษณะพหุนาม
4. ขั้นตอนวิธีเชิงพันธุกรรม
5. การจำแนกประเภทต้นไม้ตัดสินใจ
6. การปรับแต่งไฮเปอร์พารามิเตอร์
7. การแบ่งข้อมูลเพื่อวัดประสิทธิภาพแบบจำลอง
8. การวัดประสิทธิภาพแบบจำลอง
9. งานวิจัยที่เกี่ยวข้อง

ล็อกไฟล์

การบันทึกข้อมูลเหตุการณ์และกิจกรรมต่าง ๆ ที่เกิดขึ้นในระบบนั้นเป็นสิ่งสำคัญอย่างยิ่งในการดูแลความปลอดภัยของระบบคอมพิวเตอร์หรือแอปพลิเคชันต่าง ๆ ไฟล์บันทึกข้อมูลเกี่ยวกับเหตุการณ์และกิจกรรมต่าง ๆ ที่เกิดขึ้นในระบบหรือแอปพลิเคชันในระบบคอมพิวเตอร์ เรียกว่า ล็อกไฟล์ การใช้ล็อกไฟล์ในการตรวจสอบปัญหาและวิเคราะห์ความผิดปกติของเหตุการณ์และกิจกรรมของผู้ใช้บริการนั้นจะสามารถช่วยให้ผู้ดูแลระบบหรือผู้พัฒนาระบบสามารถดำเนินการแก้ไขและปรับปรุงระบบให้ดียิ่งขึ้นได้ ซึ่งนำไปสู่ระบบการตรวจจับการบุกรุกเข้าสู่ระบบคอมพิวเตอร์หรือเครือข่ายได้ (Intrusion Detection System หรือ IDS) โดยการทำงานในรูปแบบลักษณะผิดปกติ (Anomaly-based IDS หรือ AIDS) ที่เรียนรู้และสร้างตัวแบบพฤติกรรมของระบบ โดยใช้ข้อมูลจากการดูแลระบบ และจะตรวจสอบการกระทำที่มีลักษณะผิดปกติหรือไม่เป็นไปตามตัวแบบที่สร้างขึ้น การกระทำที่แตกต่างจากพฤติกรรมปกติอาจถูกพิจารณาว่าเป็นการบุกรุก โดย AIDS สามารถค้นพบการโจมตีใหม่ ๆ โดยการสแกนและตรวจสอบรูปแบบเครือข่ายที่แตกต่างจากรูปแบบการทำงานของเครือข่ายปกติอย่างมาก (Tama et al., 2019)

ประเภทของล็อกไฟล์แบ่งออกเป็น 7 ประเภท (Abdalla & Jumaa, 2022) ดังนี้

1. Application Log เป็นการบันทึกของแอปพลิเคชันที่รวมถึงเหตุการณ์ ข้อความแสดงข้อผิดพลาด หรือคำเตือนที่โปรแกรมสร้างขึ้นมา โดยบันทึกของแอปพลิเคชันจะให้ข้อมูลแก่ผู้ดูแลระบบเกี่ยวกับสถานะของแอปพลิเคชันที่ทำงานบนเซิร์ฟเวอร์

2. Web Server Log ในการใช้งานเว็บนั้นจะมีการสื่อสารระหว่างผู้ใช้งานกับเว็บ ข้อมูลการสื่อสารเหล่านี้จะถูกบันทึกไว้ในล็อกไฟล์ ที่เรียกว่า Web Log File ข้อมูลนี้จะถูกสร้างขึ้นโดยอัตโนมัติจากการโต้ตอบการทำงานของผู้ใช้งานกับเว็บนั้น รวมถึงบันทึกการเข้าถึงเซิร์ฟเวอร์ (Server Access Log) บันทึกข้อผิดพลาด (Error Log) บันทึกผู้อ้างอิง (Referrer Log) และคุกกี้ฝั่งไคลเอ็นต์ (Client-Side Cookies) ในรูปแบบของไฟล์ข้อความ บันทึกการใช้เว็บนี้จะบันทึกทุกคำขอของเว็บที่ดำเนินการโดยไคลเอ็นต์ไปยังเซิร์ฟเวอร์

3. System Log ระบบปฏิบัติการจะบันทึกเหตุการณ์ที่เกิดขึ้นในบันทึกของระบบ เช่น ความล้มเหลวของระบบ คำเตือน และข้อผิดพลาด ซึ่งในแต่ละโปรแกรมจะมีการสร้างไฟล์บันทึกที่เกี่ยวกับเซสชันของผู้ใช้งาน ที่รวมไปถึงข้อมูลเกี่ยวกับเวลาในการเข้าสู่ระบบของผู้ใช้งาน การโต้ตอบกับแอปพลิเคชัน ผลลัพธ์จากการยืนยันตัวตน และอื่น ๆ

4. Security Log การบันทึกความปลอดภัยนั้นถูกนำมาใช้เพื่อให้มีข้อมูลที่เพียงพอในการระบุการกระทำที่เป็นอันตรายหลังจากที่เคยเกิดขึ้นมาแล้วและป้องกันไม่ให้เกิดขึ้นอีก การบันทึกความปลอดภัยจะเก็บเส้นทางของข้อมูลตามช่วงที่กำหนดโดยผู้ดูแลระบบ ตัวอย่างเช่น บันทึกของไฟร์วอลล์ซึ่งประกอบด้วยข้อมูลที่เกี่ยวข้องกับแพ็กเก็ตที่ถูกกำหนดเส้นทางจากต้นทาง IP addresses ที่ถูกปฏิเสธ การบันทึกความปลอดภัยมีข้อมูลรายละเอียดที่ผู้ดูแลระบบความปลอดภัยต้องจัดการควบคุม และประเมินให้สอดคล้องกับข้อกำหนดการใช้งาน

5. Network Log การบันทึกเครือข่ายเป็นการบันทึกเหตุการณ์ต่าง ๆ ที่เกิดขึ้นบนเครือข่าย ได้แก่ การบันทึกกิจกรรมที่เป็นอันตราย การเพิ่มขึ้นของเน็ตเวิร์คทราฟฟิก การสูญเสียแพ็กเก็ต และความล่าช้าของแบนด์วิธ ทั้งนี้อาจรวบรวมข้อมูลจากอุปกรณ์เครือข่ายต่าง ๆ รวมไปถึงสวิสช์ เราเตอร์ และไฟร์วอลล์ สำหรับการบันทึกเครือข่ายนั้นสามารถนำข้อมูลไปใช้ในการตรวจสอบการโจมตีต่าง ๆ ได้

6. Audit Log การบันทึกการตรวจสอบช่วยผู้ดูแลด้านความปลอดภัยในการวิเคราะห์กิจกรรมที่เป็นอันตรายระหว่างการโจมตี ที่อยู่ต้นทางและปลายทาง การประทับเวลาและข้อมูลการเข้าสู่ระบบของผู้ใช้ จะเป็นส่วนข้อมูลที่สำคัญที่สุดของไฟล์บันทึกการตรวจสอบ

7. Virtual Machine Log การบันทึกคอมพิวเตอร์เสมือน เป็นการเก็บรายละเอียดเกี่ยวกับอินสแตนซ์ที่กำลังดำเนินการบนเครื่องคอมพิวเตอร์เสมือน (Virtual Machine หรือ VM) เช่น การกำหนดค่าเริ่มต้น การดำเนินการ และเวลาในการดำเนินการของแต่ละแอปพลิเคชัน และการย้าย

แอปพลิเคชัน ซึ่งช่วย Cloud Solution Provider (CSP) ในการระบุกิจกรรมที่เป็นอันตรายที่เกิดขึ้นขณะโจมตี

ข้อมูลบันทึก (Log Data) ถูกสร้างขึ้นโดยบริการและมีข้อความกึ่งโครงสร้างที่ผนวกเข้ากับไฟล์ที่มีนามสกุล .log ไฟล์เหล่านี้มีขนาดโตขึ้นและอาจมีขนาดใหญ่มาก (Viola et al., 2022) มีการบันทึกมากกว่า 1.4 พันล้านรายการในแต่ละวัน ทั้งนี้เพื่อใช้ในตรวจจับความผิดปกติของการดำเนินกิจกรรมและพฤติกรรมโปรไฟล์ของผู้ใช้งาน (Meena, 2022) โดย David et al. (2022) เสนอวิธีการตรวจจับความผิดปกติโดยใช้เกณฑ์เพื่อแยกแยะระหว่างไฟล์บันทึกปกติและไฟล์บันทึกที่ผิดปกติ การทดลองดำเนินการบน Hadoop Distributed File System (HDFS) ซึ่งเป็นชุดข้อมูลบันทึกที่เผยแพร่ต่อสาธารณะ ประสิทธิภาพของระบบโดยใช้ Robust Random Cut Forest (RRCF) Ritchey& Perry (2021) ใช้วิธีชุดเครื่องมือที่ใช้ Python ร่วมกับการเรียนรู้ของเครื่องที่ไม่มีผู้ดูแลเพื่อลดขนาดไฟล์บันทึกและตรวจจับพฤติกรรมที่เป็นอันตราย Ertam & Kaya (2018) จัดประเภทล็อกไฟล์ของไฟร์วอลล์โดยใช้ตัวแยกประเภท Multiclass Support Vector Machine (SVM) และประเมินประสิทธิภาพของตัวแยกประเภทโดยใช้ฟังก์ชันการเปิดใช้งานที่แตกต่างกัน

ในงานวิทยานิพนธ์นี้จึงได้นำชุดข้อมูลที่เป็นล็อกไฟล์ที่เกี่ยวกับเครือข่ายมาใช้ในการคัดเลือกคุณลักษณะที่เหมาะสมด้วยการสร้างเฟรมเวิร์คการคัดเลือกคุณลักษณะ โดยวิธีการคัดเลือกคุณลักษณะได้กล่าวในรายละเอียดในหัวข้อถัดไป

การคัดเลือกคุณลักษณะ

การคัดเลือกคุณลักษณะ คือ การลดมิติที่มีผลกระทบของข้อมูลผ่านการหาเซ็ตย่อยของคุณลักษณะซึ่งมีประสิทธิภาพในการกำหนดข้อมูล โดยจะทำการเลือกคุณลักษณะของข้อมูลที่มีความสำคัญและมีความเกี่ยวข้องในการวิเคราะห์ข้อมูลจากข้อมูลนำเข้า และทำการลบข้อมูลที่มีความซ้ำซ้อนและไม่เกี่ยวข้องออก ในการคัดเลือกคุณลักษณะที่สำคัญของล็อกไฟล์นั้น เป็นการหาคุณลักษณะที่สำคัญ ซึ่งการหาคุณลักษณะของข้อมูลที่สำคัญนั้นไม่เพียงแต่เพิ่มความเร็วในการจัดการข้อมูล แต่ยังช่วยในการปรับปรุงประสิทธิภาพของการตรวจจับข้อมูลได้ (Brandao & Georgieva, 2020) นักวิจัยต่าง ๆ ให้ความสำคัญในการที่ทำการคัดเลือกคุณลักษณะของข้อมูลและนำเสนอวิธีการที่หลายหลายในการคัดเลือกคุณลักษณะของข้อมูล

เทคนิคการคัดเลือกคุณลักษณะแบ่งออก 5 เทคนิค (Zebari et al., 2020) ดังนี้

1. วิธีการกรอง เป็นวิธีการในการประเมินความสำคัญของคุณลักษณะ โดยพิจารณาลักษณะแต่ละตัวแยกต่างหากจากข้อมูลที่เป็นอิสระ โดยใช้เครื่องมือสถิติหรือคำสั่งสร้างขึ้นเพื่อจัดอันดับหรือคะแนนคุณลักษณะ เช่น

- ค่า Information Gain วัดปริมาณข้อมูลที่คุณลักษณะให้สำหรับทำนายตัวแปรเป้าหมาย
- ค่า ไคสแควร์ ประเมินความอิสระระหว่างคุณลักษณะและตัวแปรเป้าหมายโดยใช้สถิติ ไคสแควร์

เทคนิคนี้มีประสิทธิภาพที่ดีและการประมวลผลที่มีประสิทธิภาพสูง ปรับขนาดได้ง่ายในชุดข้อมูลที่มีมิติสูง และมีประสิทธิภาพดีกว่าวิธีการห่อหุ้ม ข้อเสียหลักของวิธีนี้คือการละเลยการรวมระหว่างชุดย่อยที่ถูกเลือกและประสิทธิภาพของขั้นตอนวิธีเชิงกลุ่ม (Abd-Alsabour, 2018; Jindal & Kumar, 2017)

2. วิธีการห่อหุ้ม เป็นวิธีการที่ใช้ในการประเมินประสิทธิภาพของขั้นตอนวิธีการเรียนรู้ด้วยคุณลักษณะย่อยต่าง ๆ โดยใช้วิธีการค้นหาเพื่อสำรวจต้นทางของคุณลักษณะที่เป็นไปได้ การดำเนินการคัดเลือกคุณสมบัติตามประสิทธิภาพของขั้นตอนวิธีการเรียนรู้ โดยจะเลือกคุณลักษณะที่เหมาะสมที่สุดสำหรับขั้นตอนวิธีการทำนาย ดังนั้นจึงได้ประสิทธิภาพที่ดีขึ้นและมีความแม่นยำสูงเมื่อเทียบกับวิธีการกรอง (Jain & Singh, 2018; Jindal, & Kumar, 2017) ตัวอย่างเช่น

- วิธีการลดลักษณะอย่างละเอียด (Recursive Feature Elimination หรือ RFE) เป็นการลบลักษณะที่มีผลกระทบน้อยที่สุดต่อประสิทธิภาพของขั้นตอนวิธีการเรียนรู้ไปเรื่อย ๆ
- วิธีการขั้นตอนวิธีเชิงพันธุกรรม ใช้วิวัฒนาการในการค้นหา แทนคุณลักษณะที่เป็นโครโมโซม และใช้ตัวดำเนินการพันธุกรรม เช่น การสลับสายพันธุ์ (Crossover) และ การกลายพันธุ์ (Mutation)

ข้อเสียหลักของวิธีนี้คือความซับซ้อนในการคำนวณและการเกิด overfitting มากเกินไปเมื่อเปรียบเทียบกับวิธีการกรอง วิธีการห่อหุ้มส่วนใหญ่จะใช้ในกรณีที่เป็นหลายตัวแปร

3. วิธีการฝังตัว เป็นวิธีการคัดเลือกคุณลักษณะที่ฝังการคัดเลือกคุณลักษณะเข้าไปอยู่ในขั้นตอนวิธีการเรียนรู้และใช้คุณสมบัติของขั้นตอนวิธีการเรียนรู้ในการประเมินคุณลักษณะ วิธีการแบบฝังตัวนั้นมีประสิทธิภาพมากกว่าและจับต้องได้มากกว่าวิธีการห่อหุ้มด้วยการคำนวณในขณะที่ยังคงประสิทธิภาพที่ใกล้เคียงกัน เป็นเพราะวิธีการฝังตัวหลีกเลี่ยงการดำเนินการซ้ำของตัวแยกประเภทและการตรวจสอบคุณสมบัติย่อยทุกชุด ตัวอย่างเช่น

- Regularization-based Methods เช่น Least Absolute Shrinkage and Selection Operator หรือ LASSO และ Elastic Net ที่ใช้เทคนิคการลดค่าน้ำหนักของคุณลักษณะที่ไม่สำคัญในการสร้างตัวแบบ

- Tree-base Method เช่น Decision Tree, Random Forest, Gradient Boosting ที่สามารถปรับค่าความสำคัญของคุณลักษณะในการสร้างต้นไม้การตัดสินใจโดยใช้เกณฑ์คุณลักษณะเพื่อตัดสินใจในการแบ่งกลุ่ม

4. วิธีการผสาน เป็นวิธีการที่นำสองวิธีมารวมเข้าด้วยกัน เช่น วิธีการห่อหุ้ม และวิธีตัวกรอง โดยสองวิธีที่ทำมารวมเข้าด้วยกันจะต้องมีเกณฑ์ในการคัดเลือกคุณลักษณะเดียวกัน การรวมกันของวิธีการกรองและวิธีการห่อหุ้มเป็นวิธีการแบบผสมผสานที่พบมากที่สุด ตัวอย่างเช่น

- ReliefF เป็นการค้นหาเพื่อประเมินความสำคัญของคุณลักษณะและกำหนดน้ำหนักให้กับคุณลักษณะตามค่าที่คำนวณได้
- Boruta ใช้การสร้างต้นไม้แบบสุ่มและคุณลักษณะเทียบเท่า (Shadow Features) เพื่อกำหนดความสำคัญของคุณลักษณะโดยเปรียบเทียบกับคุณลักษณะที่สร้างขึ้นแบบสุ่ม

5. วิธีการแบบรวม เป็นวิธีการที่มีจุดมุ่งหมายเพื่อสร้างกลุ่มของส่วนย่อยของคุณลักษณะแล้วสร้างผลลัพธ์รวมจากกลุ่มนี้ขึ้นอยู่กับเทคนิคการสุ่มตัวอย่างแบบต่างๆ ซึ่งใช้วิธีการเลือกคุณลักษณะเฉพาะกับตัวอย่างย่อยต่างๆ และคุณลักษณะที่ได้รับจะถูกรวมเข้าด้วยกันเพื่อสร้างชุดย่อยที่มีเสถียรภาพมากขึ้น ตัวอย่างเช่น

- Voting Ensemble วิธีการนี้ใช้หลาย ๆ เทคนิคในการคัดเลือกคุณลักษณะเพื่อทำการโหวตหรือตัดสินใจเพื่อคัดเลือกคุณลักษณะที่สำคัญ
- Bagging Ensemble วิธีการนี้ใช้ตัวแบบหลาย ๆ เทคนิคในการคัดเลือกคุณลักษณะเพื่อสร้างตัวอย่างแบบสุ่ม (Bootstrap Sample) จากข้อมูลเพื่อทำการคัดเลือกคุณลักษณะ แล้วนำผลลัพธ์ของแต่ละตัวอย่างมาเชื่อมต่อกัน เช่น Random Forest, Extra Trees เป็นต้น
- Boosting Ensemble วิธีการนี้ใช้ตัวแบบหลาย ๆ เทคนิคในการคัดเลือกคุณลักษณะเพื่อสร้างตัวอย่างแบบเป็นลำดับ (Sequential Sample) โดยตัวแบบในแต่ละลำดับจะศึกษาและปรับปรุงตามผลลัพธ์ของตัวแบบก่อนหน้า

Zebari et al. (2020) ได้สรุปเหตุผลในการทำการคัดเลือกคุณลักษณะ โดยลดขนาดของชุดข้อมูลได้โดยไม่เสียข้อมูลที่สำคัญ อาทิ

1. ลดความซับซ้อนของตัวแบบ ตัวแบบที่มีคุณลักษณะมากเกินไปอาจทำให้เกิดการเรียนรู้ที่ไม่แม่นยำ การเลือกคุณลักษณะช่วยลดจำนวนคุณลักษณะที่ไม่เกี่ยวข้องหรือซ้ำซ้อน ทำให้ตัวแบบทำงานได้ง่ายขึ้นและมีประสิทธิภาพมากขึ้น

2. ลดเวลาและทรัพยากรในการประมวลผล การทำงานกับชุดข้อมูลที่มีคุณลักษณะมากโดยไม่จำเป็นอาจทำให้การประมวลผลช้าลงและใช้ทรัพยากรคอมพิวเตอร์มากขึ้น การเลือกคุณลักษณะช่วยลดขนาดข้อมูลที่จะถูกประมวลผล ทำให้การวิเคราะห์เร็วขึ้นและประหยัดทรัพยากร

3. ป้องกันการเรียนรู้จากข้อมูลเชิงลึก (Overfitting) ตัวแบบที่มีคุณลักษณะมากเกินไปอาจเรียนรู้รายละเอียดข้อมูลในชุดข้อมูลฝึกฝนที่ไม่สามารถแยกแยะข้อมูลทั่วไปได้ ทำให้มีประสิทธิภาพในการทำนายข้อมูลที่ใช้ในการฝึกฝนเท่านั้น การเลือกคุณลักษณะช่วยลดปัญหาที่เกิดจากการเรียนรู้เชิงลึกและทำให้ตัวแบบมีความสามารถในการทำนายที่ดีกว่า

จากประโยชน์ของการทำการคัดเลือกคุณลักษณะ ผู้วิจัยได้นำวิธีการการคัดเลือกคุณลักษณะด้วยวิธีการกรองที่ใช้ค่าทางสถิติต่าง ๆ มาใช้เป็นพื้นฐานในการคัดเลือกคุณลักษณะของชุดข้อมูล โดยใช้ค่าทางสถิติในการเปรียบเทียบของแต่ละชุดข้อมูล ดังนี้

1. ไคสแควร์ เป็นสถิติที่กำหนดระดับความเป็นอิสระระหว่างคุณลักษณะที่ a_i และคลาสที่ y_j และเปรียบเทียบการกระจายของไคสแควร์ด้วยค่าระดับ degree of freedom เท่ากับ 1 ดังนั้นสถิติไคสแควร์จะได้ดังสมการ (1) (Thaseen et al., 2019)

$$\chi^2(a_i, y_j) = \frac{N \cdot (TZ - YX)^2}{(T+X)(T+Z)(X+Z)(Y+Z)} \quad (1)$$

โดยที่ T คือ ความถี่ของคุณลักษณะที่ a_i คลาสที่ y_j ของชุดข้อมูล

X คือ ความถี่ของ a_i ที่ไม่ปรากฏ y_j

Y คือ ความถี่ของ y_j ที่ไม่ปรากฏ a_i

Z คือ ในกรณีที่ไม่ใช่ทั้ง y_j หรือ a_i ที่ปรากฏพร้อมกันในชุดข้อมูล

N คือ จำนวนเรคคอร์ดรวม

2. การวิเคราะห์ความแปรปรวน เป็นวิธีการทางสถิติที่ใช้สำหรับการเปรียบเทียบวิธีการที่เป็นอิสระต่อกัน วิธีการ การวิเคราะห์ความแปรปรวน จัดอันดับคุณลักษณะโดยการคำนวณอัตราส่วนของความแปรปรวนระหว่างกลุ่มและภายในกลุ่ม (Nasiri & Alavi, 2022) ดังสมการ (2)

$$F = \frac{MSB}{MSE} \quad (2)$$

โดยที่ F คือ ความแปรปรวน

MSB คือ ผลรวมกำลังสองเฉลี่ยระหว่างกลุ่ม

MSE คือ ผลรวมกำลังสองเฉลี่ยภายในกลุ่ม

3. สารสนเทศร่วม (Mutual Information) แนวคิดของข้อมูลร่วมกันเพื่ออธิบายข้อมูลทั่วไประหว่างตัวแปร เช่น ระดับของการลดความไม่แน่นอนของตัวแปรหนึ่งเมื่อทราบอีกตัวแปรหนึ่ง ให้ตัวแปรสุ่มสองตัวสามารถกำหนดข้อมูลร่วมกันได้ดังสมการ (3) (Song et al., 2021)

$$I(X,Y) = \sum_y \sum_x p(x,y) \lg \frac{p(x,y)}{p(x)p(y)} \quad (3)$$

โดยที่ x คือ ตัวอย่างของตัวแปร X

y คือ ตัวอย่างของตัวแปร Y

$p(x)$ คือ ความน่าจะเป็นของ X

$p(y)$ คือ ความน่าจะเป็นของ Y

$p(x,y)$ คือ ความน่าจะเป็นของ X และ Y ร่วมกัน

4. ความสำคัญของคุณลักษณะ

กำหนดให้ f แทนคุณลักษณะบางคุณลักษณะของเซ็ตของคุณลักษณะทั้งหมด การมีส่วนร่วมของ f ไปยังคลาส c จะอธิบายโดยความสำคัญของคุณลักษณะ

ความสำคัญของ f แทนด้วย $GI(f)$ ชุดของข้อมูล D สามารถวัดความแน่นอนของโหนดได้โดยค่า Gini index in CART ดังสมการ (4), (5)

$$Gini(p_1, \dots, p_y) = \sum_{j=1}^{y} p_j(1 - p_j) \quad (4)$$

โดยที่ $p_1, \dots, p_y$ คือ ความน่าจะเป็นของประเภท 1, ..., y

เกณฑ์สำหรับ CART คือการลด Gini ให้ได้สูงสุด เมื่อแยกโหนดตามคุณลักษณะ f การลดลงของ Gini ของคุณลักษณะ f จะถูกกำหนดเป็น

$$GD(f) = Gini(D) - \sum_v \frac{|D^v|}{|D|} Gini(D^v) \quad (5)$$

โดยที่ D^v คือ ชุดข้อมูลที่แบ่งแยกเป็นโหนดลูก

การสร้างคุณลักษณะพหุนาม

การสร้างคุณลักษณะพหุนาม เป็นเทคนิคที่ใช้ในเทคนิคการเลือกคุณลักษณะประเภทของวิศวกรรมคุณลักษณะ (Feature Engineering) ซึ่งเป็นขั้นตอนการสร้างคุณลักษณะใหม่จากคุณลักษณะเดิมโดยใช้การสร้างพหุนามของคุณลักษณะเดิมขึ้นมา ในกระบวนการนี้จะสร้าง

คุณลักษณะใหม่โดยการเพิ่มพื้นที่และความซับซ้อนให้กับข้อมูลเดิม ถ้าคุณลักษณะเดิมที่เป็นเชิงเส้นจะสามารถแปลงคุณลักษณะนั้นเป็นพหุนาม โดยการเพิ่มเป็นสมการพหุนามเชิงเส้นที่มีกำลังสูงขึ้น โดยปกติจะใช้ฟังก์ชันของคุณลักษณะที่เป็นเชิงเส้นอย่างเดียวในการสร้างพหุนาม เช่น กำลังสอง หรือกำลังสามของคุณลักษณะเดิม

ตัวอย่างของชุดข้อมูลที่ประกอบไปด้วยเวกเตอร์คุณลักษณะ 2 คุณลักษณะ (Oswald et al., 2021)

$$X = [x_1, x_2] \tag{6}$$

เมื่อคุณลักษณะพหุนามเป็นเวกเตอร์กำลัง 2 จะได้

$$X = [1, x_1, x_2, x_1^2, x_1x_2, x_2^2] \tag{7}$$

และสำหรับพหุนามระดับสูงขึ้นไป ควรสังเกตว่าเวกเตอร์คุณลักษณะพหุนามใหม่ซึ่งประกอบด้วย

1. ค่าปริมาณความเอนเอียง (Bias)
2. ตัวแปรดั้งเดิมทั้งหมดยกกำลังขึ้นตามระดับที่กำหนด
3. ตัวแปรปฏิสัมพันธ์

การเพิ่มคุณลักษณะใหม่ในรูปแบบพหุนามช่วยให้ตัวแบบสามารถจับความสัมพันธ์ที่ซับซ้อนของข้อมูลได้ดีขึ้น เช่น การแสดงแบบจำลองที่มีความสัมพันธ์เชิงกำลัง (non-linear relationship) ระหว่างตัวแปร และให้ตัวแบบสามารถปรับผลกระทบของคุณลักษณะในรูปแบบพหุนามได้ การสร้างคุณลักษณะพหุนามเป็นเทคนิคที่นิยมใช้ในงานเชิงพยากรณ์ (predictive modeling) เช่น Regression และ Polynomial Regression แต่อาจใช้ได้ในรูปแบบอื่น ๆ ของตัวแบบที่ต้องการความซับซ้อนในข้อมูลได้

ขั้นตอนวิธีเชิงพันธุกรรม

ขั้นตอนวิธีทางพันธุกรรม เป็นขั้นตอนวิธีในการปรับให้เหมาะสมที่ได้รับแรงบันดาลใจจากการคัดเลือกโดยธรรมชาติ (Katoch, 2021) เป็นขั้นตอนวิธีในการค้นหาและหาค่าที่เหมาะสมโดยใช้กระบวนการสุ่มของกลุ่มวิธีการ (Zhu et al., 2022) อิงหลักการคัดเลือกตามธรรมชาติของดาร์วิน (Darwin's theory of natural selection) ซึ่งแนวคิดนี้ถูกนำมาประยุกต์ใช้ในการแก้ปัญหาโดย

John Holland และได้ใช้ชื่อว่า “วิธีเชิงพันธุกรรม” แต่ก็ยังไม่ได้รับความนิยม จนกระทั่ง David Goldberg ได้นำมาตีพิมพ์เป็นหนังสือโดยอธิบายรายละเอียดต่างๆ ของ ขั้นตอนวิธีเชิงพันธุกรรม ตลอดจนวิธีการนำไปประยุกต์ใช้ จนทำให้ ขั้นตอนวิธีเชิงพันธุกรรม เป็นที่นิยมของนักวิจัยในการนำมาใช้ในการวิจัยและมีลักษณะที่แตกต่างกันไปตามงานวิจัยและแนวทางการพัฒนาของนักวิจัยนั้นๆ ซึ่งโดยทั่วไปจะมีโครงสร้างมาจาก ขั้นตอนวิธีเชิงพันธุกรรม ตัวต้นแบบที่เรียกว่า “ขั้นตอนวิธีทางพันธุกรรมอย่างง่าย” ซึ่งหลักการทำงานหลัก ๆ ประกอบไปด้วย 5 ขั้นตอน

1. การเข้ารหัสโครโมโซม เป็นการสุ่มค่าคำตอบต่าง ๆ ที่อยู่ภายในขอบเขตแล้วนำไปเข้ารหัส เพื่อทำให้อยู่ในรูปแบบของโครโมโซม ซึ่งโดยทั่วไปแล้วจะอยู่ในรูปของตัวเลขฐานสอง และเรียกกลุ่มของโครโมโซมที่ได้จากการสุ่มนี้ว่าประชากร

2. กระบวนการทางพันธุกรรม เป็นวิธีการที่จะนำเอาโครโมโซมที่ได้จากการสุ่มจากขั้นตอนที่แล้วไปกระทำตามขั้นตอนทางพันธุกรรม ซึ่งประกอบด้วย 2 กระบวนการ คือ

2.1 การสลับสายพันธุ์ เป็นการนำโครโมโซมที่ได้จากการสุ่มมาจับคู่เพื่อเป็นโครโมโซมพ่อและโครโมโซมแม่ แล้วทำการสุ่มแลกเปลี่ยนพันธุกรรมเพื่อให้ได้โครโมโซมใหม่ที่แตกต่างออกไปจากโครโมโซมพ่อและโครโมโซมแม่ ซึ่งวิธีที่ง่ายที่สุดคือการสลับสายพันธ์แบบจุดเดียว

2.2 การกลายพันธุ์ เป็นการนำโครโมโซมที่ผ่านการสลับสายพันธุ์แล้วมาทำการสุ่มแล้วเปลี่ยนค่าของยีนส์เพื่อให้เกิดโครโมโซมใหม่ขึ้นมาอีกหนึ่งชุด โดยโครโมโซมที่เกิดขึ้นมาใหม่นี้จะเรียกว่า “โครโมโซมลูก ”

3. การคำนวณค่าความเหมาะสม เมื่อโครโมโซมผ่านกระบวนการทางพันธุกรรมแล้วจะถูกประเมินค่าความเหมาะสม โดยฟังก์ชั่นวัตถุประสงค์ที่ใช้ในการประเมินสมรรถภาพ ของโอกาสการอยู่รอดของแต่ละโครโมโซม

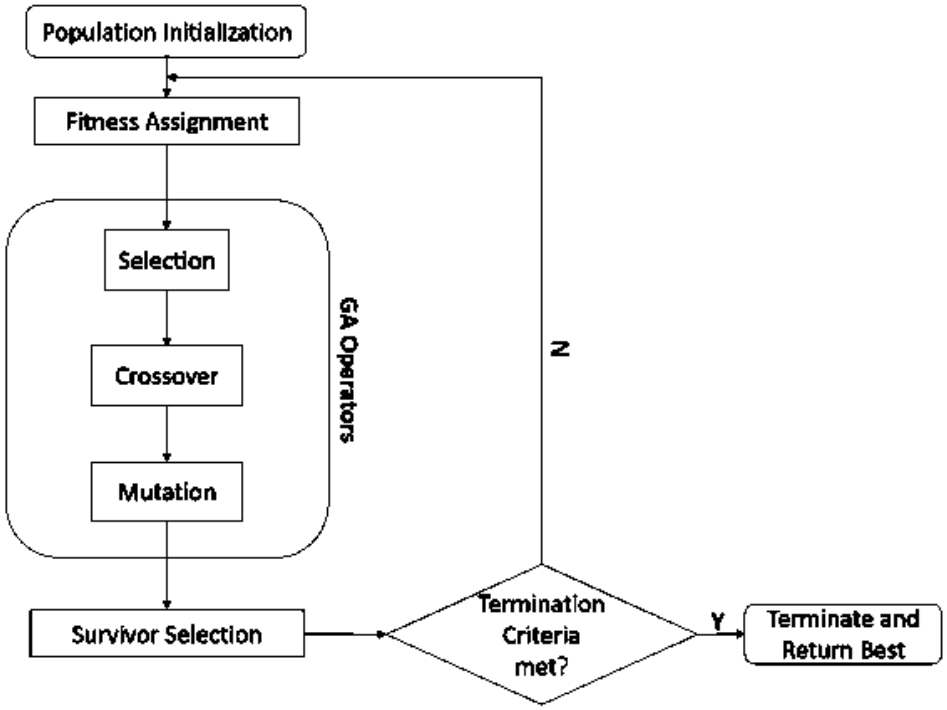
4. การคัดเลือกสายพันธุ์ กลไกลการคัดเลือกสายพันธุ์ของ ขั้นตอนวิธีเชิงพันธุกรรมนั้นขึ้นอยู่กับค่าของความเหมาะสมในการอยู่รอด ถ้าโครโมโซมที่มีค่าของความเหมาะสมของการอยู่รอดสูงจะทำให้โครโมโซมนั้นมีโอกาศที่จะถูกเลือกไปเป็นประชากรในรุ่นถัดต่อไปสูง แต่ถ้าโครโมโซมที่มีค่าของความเหมาะสมของการอยู่รอดต่ำโอกาศของการที่โครโมโซมนั้นจะถูกเลือกเพื่อไปเป็นประชากรในรุ่นถัดไปก็จะน้อยตามไปด้วย ซึ่งวิธีการที่นิยมนำมาใช้เพื่อการคัดเลือกสายพันธุ์นั้น หลัก ๆ มีอยู่ 3 กระบวนการ คือ

4.1 การคัดเลือกแบบการจัดอันดับ เป็นการนำค่าความเหมาะสมในการอยู่รอดของแต่ละโครโมโซมมาทำการเรียงลำดับแล้วเลือกตามความเหมาะสม

4.2 การคัดเลือกแบบการแข่งขัน คือ การสุ่มจับคู่เปรียบเทียบจากกลุ่มประชากรและคัดเลือกผู้ชนะจากการเปรียบเทียบนั้น

4.3 การคัดเลือกแบบวงล้อรูเล็ต คือ การสุ่มเลือกด้วยการกำหนดความน่าจะเป็นในการถูกคัดเลือกตามสัดส่วนของคะแนนความเหมาะสมของประชากรจากผลรวมคะแนนทั้งหมด

5. การตรวจสอบเงื่อนไขการสิ้นสุดการทำงาน การสิ้นสุดการทำงานของ ขั้นตอนวิธีเชิงพันธุกรรมนั้น ขึ้นอยู่กับการกำหนดตัวเลขของจำนวนประชากร (Number of generation) เมื่อมีการทำงานจนครบตามจำนวนประชากรที่ได้กำหนดไว้แล้วให้ทำการหยุดทำงาน หรืออาจจะเป็นตามที่ผู้ใช้กำหนดไว้ว่าถ้าหากคำตอบที่ออกมาไม่มีการเปลี่ยนแปลงในหลาย ๆ รอบให้หยุดการทำงาน



ภาพ 1 ขั้นตอนการทำงานของกระบวนการทางพันธุกรรม

ที่มา: packt, 2023

แสดงให้เห็นการนำขั้นตอนวิธีทางพันธุกรรมไปประยุกต์ใช้โดยการทำงานเริ่มที่ การสุ่มสร้างโครโมโซมต้นแบบขึ้นมา เพื่อไปใช้ในการคำนวณหาค่าของความเหมาะสมในขั้นตอนต่อไป จะเห็นได้ว่ามีการตรวจสอบการทำงาน โดยที่จะมีการให้ตรวจสอบเงื่อนไขการทำงานว่าถ้าหากค่าความเหมาะสมใกล้เคียงต่อความต้องการแล้ว จะเป็นการหยุดทำงานแต่ถ้าหากค่าความเหมาะสมของโครโมโซมยังไม่ใกล้เคียงให้ทำตามขั้นตอนทางพันธุกรรมเพื่อให้ได้กลุ่มของประชากรใหม่เพื่อที่จะนำไปใช้ในการคำนวณหาค่าความเหมาะสมตามความต้องการของผู้ใช้ ซึ่งภายในขั้นตอนทาง

พันธุกรรมจะประกอบไปด้วย การคัดเลือกโครโมโซมพ่อและโครโมโซมแม่ การแลกเปลี่ยนพันธุกรรม การกลายพันธุ์ จนกระทั่งถึงการตรวจสอบเงื่อนไขในการหยุดทำงาน ซึ่งในที่นี้ได้กำหนดเป็นจำนวนขนาดของประชากร โดยที่ถ้าหากขนาดของประชากรครบตามที่กำหนดแล้วจะทำการนำประชากรเข้าการตรวจสอบค่าความเหมาะสมเพื่อให้ได้มาซึ่งคำตอบ และการทำงานนี้จะจบลงก็ต่อเมื่อมีโครโมโซมที่มีค่าของความเหมาะสมใกล้เคียงต่อความต้องการของผู้ใช้มากที่สุด แต่ถ้ายังไม่ได้โครโมโซมที่มีค่าความต้องการก็จะทำไปเรื่อย ๆ ตามขั้นตอนของการทำงาน

ขั้นตอนวิธีเชิงพันธุกรรมเป็นวิธีการที่อยู่ในการคัดเลือกคุณลักษณะประเภทวิธีการห่อหุ้ม (Aalaei et al., 2016) การนำขั้นตอนวิธีเชิงพันธุกรรมไปใช้ในการคัดเลือกคุณลักษณะนั้นมีนักวิจัยได้นำมาใช้ในการคัดเลือกคุณลักษณะของชุดข้อมูลเพื่อให้ได้ประสิทธิภาพในการวิเคราะห์ข้อมูลที่สูงขึ้น เช่น Aalaei et al. (2016) ใช้ขั้นตอนวิธีทางพันธุกรรมในการคัดเลือกคุณลักษณะของชุดข้อมูลมะเร็งเต้านมของรัฐวิสคอนซิน และใช้วิธีการสร้างตัวแบบหลาย ๆ ตัวแบบมาทำการเปรียบเทียบประสิทธิภาพในการทำนายหลังจากที่ได้ทำการคัดเลือกคุณลักษณะแล้ว ซึ่งผลการวิจัยพบว่าการคัดเลือกคุณลักษณะนั้นสามารถปรับปรุงค่าความถูกต้องได้ Khotimah et al. (2020) ได้ใช้ขั้นตอนวิธีทางพันธุกรรมในการเป็นวิธีการจำแนกประเภทตามการคัดเลือกคุณลักษณะที่ทำให้เกิดปัญหาในการคำนวณ เช่น ขนาดที่ลดลง ความไม่แน่นอนของข้อมูล และชุดข้อมูลที่ไม่สมดุลกับคลาสต่าง ๆ แล้วนำมาใช้ในการทำงานร่วมกับ Naïve Bayes ซึ่งให้ค่าความแม่นยำที่สูงขึ้น Amini & Hu (2021) ได้ใช้วิธีการคัดเลือกคุณลักษณะแบบสองชั้นด้วยการใช้วิธีการห่อหุ้มและวิธีการแบบฝังตัวในการสร้างชุดข้อมูลย่อยที่เหมาะสมกับตัวทำนาย โดยชั้นแรกใช้ขั้นตอนวิธีทางพันธุกรรมในการลดจำนวนคุณลักษณะและข้อผิดพลาดในการทำนาย และชั้นที่สองใช้ Elastic Net (EN) ในการกำจัดตัวทำนานที่ซ้ำซ้อนและไม่เกี่ยวข้องใด ๆ ที่ยังคงเหลืออยู่หลังจากการใช้ขั้นตอนวิธีทางพันธุกรรม ซึ่งเป็นวิธีการฝังตัว เป็นต้น

ในงานวิจัยนี้จึงได้มีใช้ขั้นตอนวิธีทางพันธุกรรมที่เป็นวิธีการคัดเลือกคุณลักษณะแบบห่อหุ้มมาทำงานร่วมกับการสร้างคุณลักษณะพหุนามที่เป็นเทคนิคในการคัดเลือกทางวิศวกรรมคุณลักษณะมาใช้ในการคัดเลือกคุณลักษณะและเมื่อได้คุณลักษณะและไฮเปอร์พารามิเตอร์ที่เหมาะสมแล้วนั้น จะได้นำคุณลักษณะไปสร้างตัวแบบการพยากรณ์เพื่อทำการหาประสิทธิภาพความถูกต้องของตัวแบบด้วยการจำแนกประเภทแบบการสุ่มป่าไม้

การจำแนกประเภทด้วยต้นไม้ตัดสินใจ

การจำแนกประเภทด้วยต้นไม้ตัดสินใจ เป็นตัวแบบต้นไม้ตัดสินใจที่จัดเป็น Supervised Machine Learning Algorithm ซึ่งในการศึกษาที่ผ่านมาพบว่ามีประสิทธิภาพด้านการตรวจสอบการ

โจมตีที่เหนือกว่าการทำงานด้วยตัวจำแนกประเภทแบบ Rule-based และตัวจำแนกประเภทแบบ function-based (Barbará et al., 2001)

ต้นไม้ช่วยตัดสินใจ (Decision Tree) การสร้างต้นไม้ช่วยตัดสินใจถูกพัฒนาขึ้นโดย Quinlan, J. R., (1986) เป็นอัลกอริทึมที่นิยมนำมาใช้กันอย่างแพร่หลายของงานด้านการจำแนกประเภทข้อมูล ซึ่งโครงสร้างของต้นไม้ช่วยตัดสินใจประกอบด้วยโหนดราก (Root Node) และโหนดใบ (Leaf Node) สำหรับการสร้างต้นไม้ช่วยตัดสินใจนั้นจะขึ้นอยู่กับการคำนวณค่าความสัมพันธ์ระหว่างแอททริบิวต์กับคลาสคำตอบ (Class Label) ซึ่งสามารถคำนวณได้จากค่า Entropy และค่า Information Gain (IG) โดยแอททริบิวต์ที่มีความสัมพันธ์กับคลาสคำตอบมากที่สุดจะถูกเลือกให้เป็นโหนดราก (Kaur et al., 2015) หรือโหนดเริ่มต้นของเหตุการณ์ ชนิดของต้นไม้ช่วยตัดสินใจมีหลายชนิด สำหรับอัลกอริทึมที่นิยมนำมาใช้การจำแนกประเภทข้อมูลได้แก่ ID3 และ C4.5 หรือ J48

ต้นไม้ตัดสินใจอัลกอริทึม ID3 จะใช้ค่าของ Entropy เป็นตัววัดความแตกต่างของข้อมูล ถ้าข้อมูลมีความแตกต่างกันมาก Entropy จะมีค่าสูง แต่ถ้าข้อมูลมีความแตกต่างกันน้อย Entropy จะมีค่าต่ำ สำหรับการหาค่า Entropy นั้นคำนวณได้จากสมการ (8)

$$E(S) = -\sum_{i=1}^{n} P(X_i) log_2 P(X_i) \quad (8)$$

โดยที่

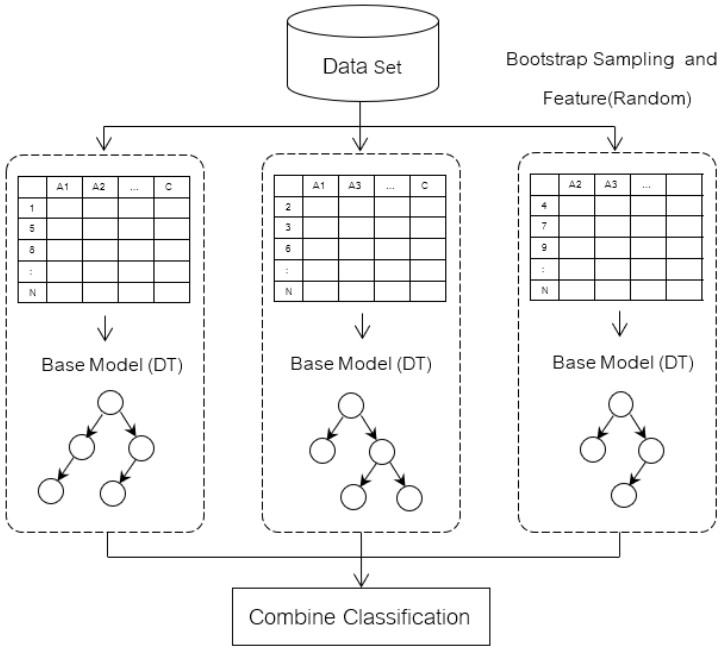
$P(X_i)$ คือ ความน่าจะเป็นที่เหตุการณ์ X_i เกิดขึ้นจากเหตุการณ์ทั้งหมด n เหตุการณ์

X_i คือ เหตุการณ์ที่จะเกิดขึ้น

n คือ จำนวนของเหตุการณ์

สำหรับในงานวิทยานิพนธ์นี้ เมื่อได้คุณลักษณะที่ได้ทำการคัดเลือกคุณลักษณะแล้วนั้น จะได้นำคุณลักษณะไปสร้างตัวแบบการพยากรณ์เพื่อทำการหาประสิทธิภาพความถูกต้องของตัวแบบด้วยการจำแนกประเภทแบบแรนดอมฟอร์เรสต์ ซึ่งการจำแนกประเภทแบบแรนดอมฟอร์เรสต์ เป็นวิธีการแบบ Bagging Ensemble โดยประกอบไปด้วยต้นไม้ช่วยตัดสินใจสามารถใช้ในการจำแนกประเภทหรือการถดถอย (Regression) ในการจำแนกประเภท การคาดคะเนจะขึ้นอยู่กับคะแนนเสียงส่วนใหญ่ของการคาดคะเนโดยใช้ต้นไม้ช่วยตัดสินใจ แต่ในกรณีของการถดถอยผลลัพธ์จะเป็นค่าเฉลี่ยของผลลัพธ์แผนผังต้นไม้ ในระหว่างขั้นตอนการฝึกสอน ข้อมูลฝึกสอน T_i จะถูกสร้างขึ้นสำหรับต้นไม้แต่ละต้น โดยอิงจากตัวอย่างในข้อมูลฝึกสอนดั้งเดิม T และเพื่อสร้างการแบ่งต้นไม้แต่ละส่วน

คุณลักษณะ m จะถูกสุ่มเลือกจากนั้นจึงทำการวิเคราะห์ด้วยการวัดในการแบ่งส่วนต้นไม้ เนื่องจากการสุ่มนี้ จึงมีการสร้างต้นไม้หลายต้น ซึ่งโดยปกติแล้วจะส่งผลให้ประสิทธิภาพการทำนายดีขึ้น แบบจำลองแรนดอมฟอร์เรสต์มีข้อดีหลายประการที่ดีกว่าวิธีการเรียนรู้ของเครื่องที่ใช้โดยทั่วไป รวมถึงระยะเวลาในการสร้างตัวแบบที่ต่ำที่สุด จัดการชุดข้อมูลที่ไม่สอดคล้องกัน กลไกการจัดหมวดหมู่สำหรับคุณลักษณะแบบฝังตัว และเมตริกภายในสำหรับกำหนดผลกระทบของคุณลักษณะ (Alduailij et al., 2022) โดยแสดงโครงสร้างวิธีการทำงานแบบแรนดอมฟอร์เรสต์ ได้ดังภาพ



ภาพ 2 โครงสร้างวิธีการทำงานแบบแรนดอมฟอร์เรสต์

การปรับแต่งไฮเปอร์พารามิเตอร์(Hyperparameter Tuning)

ในการควบคุมการทำงานของแมชชีนเลิร์นนิ่งอัลกอริทึม สำหรับการฝึกสอนโมเดลเกี่ยวกับข้อกำหนดของเครื่องที่ใช้ในการสร้างโมเดล ส่งผลโดยตรงกับ ค่าความแม่นยำของโมเดล การสร้างโมเดลการเรียนรู้ของเครื่องประกอบไปด้วยพารามิเตอร์ที่แตกต่างกัน 2 ชนิด ได้แก่

1. โมเดลพารามิเตอร์คือ พารามิเตอร์ที่ได้มาระหว่างขั้นตอนการเรียนรู้ข้อมูลของโมเดล เช่น ค่า Weights ที่ใช้ใน Neural Network หรือค่า Coefficients ที่ได้จากการทำ Linear Regression เป็นต้น

2. ไฮเปอร์พารามิเตอร์ คือ พารามิเตอร์ต่าง ๆ ที่สามารถกำหนดค่าเองได้ ที่โมเดลจะทำการเรียนรู้ เช่น ค่า Learning Rate ที่ใช้ในการควบคุมว่าใน 1 Step ของการเรียนรู้จะปรับค่า Weights

ของ Neural Network หรือการกำหนดค่า n_estimators ซึ่งกำหนดจำนวนต้นไม้สำหรับการสร้างโมเดล Random Forest เป็นต้น

Model Parameters กำหนดจะใช้ข้อมูลในการเรียนรู้ เพื่อให้ได้ผลลัพธ์ที่ต้องการโดยจะได้มาระหว่างการเรียนรู้ของโมเดล แต่ Hyperparameters ใช้สำหรับกำหนดโครงสร้างของโมเดลตั้งแต่ต้น ดังนั้นการทำ การปรับไฮเปอร์พารามิเตอร์ จะถูกเรียกว่า ไฮเปอร์พารามิเตอร์ออฟติมัย นับได้ว่าเป็น ปัญหาการหาค่าที่เหมาะสมที่สุด รูปแบบหนึ่ง เนื่องจากต้องการหาว่า เซต ของไฮเปอร์พารามิเตอร์ ที่เหมาะสมสำหรับโมเดลประเภทนั้น ๆ ที่จะส่งผลให้โมเดลมีความแม่นยำ ที่สูง หรือต้องการลดค่า Loss ให้มีค่าต่ำที่สุด

ในปัจจุบันมีเทคนิคมากมายที่ได้ถูกคิดค้นมาเพื่อใช้สำหรับการปรับแต่งไฮเปอร์พารามิเตอร์ จะแบ่งออกเป็น 2 ประเภทใหญ่ ๆ ได้แก่

1. การปรับค่าชุดไฮเปอร์พารามิเตอร์โดยตนเอง คือ วิธีการปรับค่าชุด ไฮเปอร์พารามิเตอร์ โดยจะเป็นการปรับค่าด้วยการเทียบผลของโมเดลทุกการผสมผสานของ ไฮเปอร์พารามิเตอร์ หรือปรับค่าด้วยตนเองไปเรื่อย ๆ จนกว่าจะเจอชุดของ ไฮเปอร์พารามิเตอร์ ที่ส่งผลให้โมเดลบรรลุผลตามที่คาดหมายไว้
2. การปรับค่าชุดไฮเปอร์พารามิเตอร์ที่เหมาะสมโดยอัตโนมัติ คือ การปรับค่าชุดไฮเปอร์พารามิเตอร์ ที่เหมาะสมโดยอัตโนมัติด้วยอัลกอริทึมชนิดต่าง ๆ ที่ถูกออกแบบมาเพื่องานประเภทนี้

เพื่อแสดงให้ถึงผลของการทำปัญหาการหาค่าที่เหมาะสมที่สุด ในแบบต่าง ๆ ของเทคนิคประเภท การปรับค่าชุดไฮเปอร์พารามิเตอร์โดยตนเองจะใช้ข้อมูล Titanic จาก Kaggle เพื่อเปรียบเทียบประสิทธิภาพและเวลาในการคำนวณ (Computational Time) ของโมเดลการจำแนกประเภทด้วย แรนดอมฟอร์เรสต์ เพื่อทำนายว่าผู้โดยสารในเหตุการณ์เรือไททานิคล่มจะรอดชีวิตหรือไม่รอดชีวิต (Binary Classification) แรนดอมฟอร์เรสต์ คือ หนึ่งในโมเดลการเรียนรู้แบบมีผู้สอน (Supervised Learning Model) โมเดลนี้สร้างจาก ดิซิชั่นทรี หลาย ๆ โมเดล (ต้นไม้) ตั้งแต่ 10 ต้นจนถึงมากกว่า 1000 ต้น มารวมกัน โดยข้อมูลที่จะแบ่งออกเป็นหลาย ๆ ชุดไม่ซ้ำกันเพื่อนำไปสร้างโมเดลเพื่อทำนายคลาส (Prediction) หลังจากที่ดิซิชั่นทรี ทุกโมเดลทำนายคลาสแล้ว คลาสไหนที่ที่มีคะแนนโหวตมากที่สุดจะกลายเป็นการทำนายโดยรวมของโมเดลแรนดอมฟอร์เรสต์ วิธีการค้นหาไฮเปอร์พารามิเตอร์ประเภทนี้มี 3 วิธี ได้แก่

1. Manual Search

2. Grid Search

3. Random Search

โดยผู้วิจัยจะขออธิบายรายละเอียดในแต่ละวิธีการดังนี้

1. Grid Manual Search

สำหรับวิธีนี้ จะเลือกค่าไฮเปอร์พารามิเตอร์ ของโมเดลจากประสบการณ์และความคิดเห็นส่วนบุคคล โดยจะทำการสร้างโมเดลขึ้นมาจากค่าที่เลือกและวัดความแม่นยำไปเรื่อย ๆ จนกว่าจะได้ค่าความแม่นยำที่พึงพอใจ พารามิเตอร์หลักที่ใช้สำหรับ แรนดอมฟอร์เรสต์ได้แก่

- Criterion (ค่าตั้งต้น = gini) คือ ฟังก์ชันที่ใช้ในการวัดประสิทธิภาพของการแยกโหนดของซิชั่นทรี สามารถเลือกได้ระหว่าง gini (Gini Impurity) หรือ Entropy (Information Gain)
- max_depth (ค่าตั้งต้น = None) คือ ค่าความลึกของต้นไม้ แต่ละต้นในแรนดอมฟอร์เรสต์ ยิ่งต้นไม้มีความลึกมากจะสามารถแยกข้อมูลได้ละเอียดมากขึ้น
- max_features (ค่าตั้งต้น = auto) คือ ค่าที่กำหนดจำนวนของคุณลักษณะที่ ซิชั่นทรีแต่ละต้นจะสามารถใช้ในการสร้างโมเดล
- min_samples_leaf (ค่าตั้งต้น = 1) คือ จำนวนข้อมูลขั้นต่ำใน Leaf Node ของแต่ละซิชั่นทรี ถ้าจำนวนข้อมูลต่ำกว่าค่านี้จะหยุดการแยกโหนด
- min_samples_split (ค่าตั้งต้น = 2) คือ จำนวนขั้นต่ำที่จำเป็นในโหนดเพื่อทำให้เกิดการแยกโหนด
- n_estimators (ค่าตั้งต้น = 100) คือ จำนวน Decision Tree ที่จะใช้ใน แรนดอมฟอร์เรสต์ โดยปกติแล้วยิ่งจำนวนสูงยิ่งส่งผลให้ประสิทธิภาพของโมเดลดียิ่งขึ้น แต่จะทำให้เวลาที่ใช้ในการสร้างโมเดลนานขึ้นเช่นกัน

2. Grid Search

Grid Search หรือการค้นหาแบบกริด เป็นเทคนิคที่ใช้ในการหาค่าไฮเปอร์พารามิเตอร์ ที่เข้าใจง่ายและตรงไปตรงมา ด้วยการลองใช้พารามิเตอร์ที่กำหนดไว้ล่วงหน้าทุกชุด และประเมินประสิทธิภาพหรือความแม่นยำของโมเดลแต่ละชุด จะเป็นการลองสร้างโมเดลจากค่าของ ไฮเปอร์พารามิเตอร์ ทุกชุด รูปแบบของการทำงานจะคล้ายกริด โดยค่าทั้งหมดจะอยู่ในรูปของเมทริกซ์ พารามิเตอร์แต่ละชุดจะถูกนำมาพิจารณาและสังเกตความถูกต้อง เมื่อชุดของ ไฮเปอร์พารามิเตอร์ทั้งหมดได้รับการประเมินแล้ว โมเดลที่มีชุดพารามิเตอร์ที่ให้ความแม่นยำสูงสุดจะถือว่าดีที่สุด สำหรับ

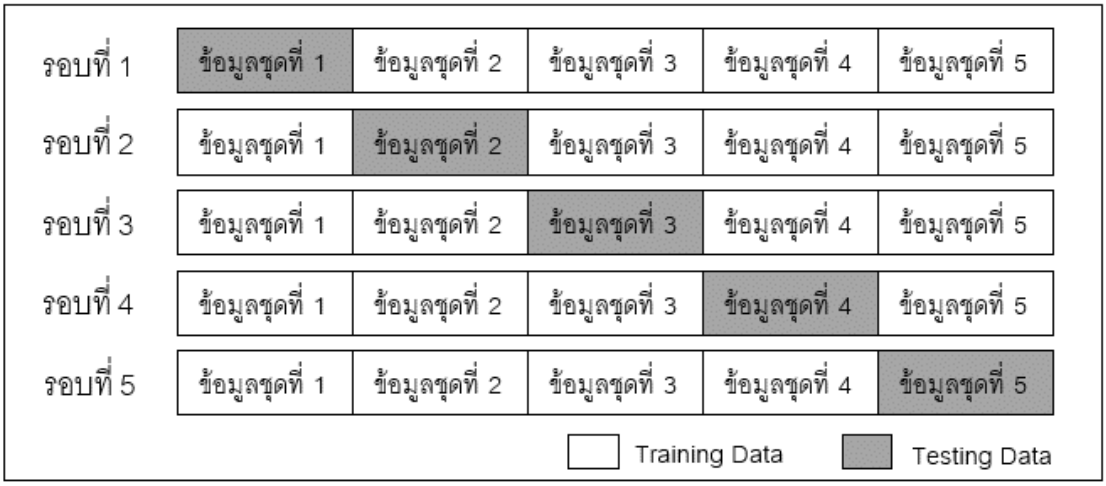
ตัวอย่างการเขียนโค้ดของ Grid Search จะเพิ่มการทำ Cross-Validation สำหรับการสร้างโมเดลเข้าไปด้วย

3. Random Search

วิธีการทำงานของ Random Search คล้ายคลึงกับการทำ Grid Search แต่แทนที่จะลองใช้พารามิเตอร์ที่กำหนดไว้ล่วงหน้าในกริดทุกชุด RandomSearch จะทำการสุ่มเลือกค่าพารามิเตอร์จากกริดที่สร้างขึ้น ดังนั้นการทำ Random Search จะไม่รับประกันว่าจะได้โมเดลที่มีประสิทธิภาพที่สุดเหมือนกับ Grid Search แต่วิธีมีประสิทธิภาพสูงในการใช้งานจริงเนื่องจากใช้เวลาในการสร้างโมเดลที่น้อยมาก ในการเขียนโค้ดเพื่อทำ Random Search จะใช้ Library จาก ไซคิทเลิร์นที่เรียกว่า RandomizedSearchCV() โดยจะแบ่ง ข้อมูลฝึกสอน ของเป็น 4 Folds (cv = 4) และจะให้โมเดลทำการสุ่มค่า ไฮเปอร์พารามิเตอร์ ออกมา 100 ชุด (n_iter = 100) เพื่อหาชุดของค่าพารามิเตอร์ที่ดีที่สุด

การแบ่งข้อมูลเพื่อวัดประสิทธิภาพตัวแบบ

การแบ่งข้อมูลเพื่อใช้สำหรับวัดประสิทธิภาพตัวแบบด้วยการจำแนกประเภทแบบการสุ่มป่าไม้นั้นมีหลายวิธีการ สำหรับในวิทยานิพนธ์นี้ ได้ทำการแบ่งข้อมูลด้วยวิธีการ Split Test ซึ่งเป็นการแบ่งข้อมูลด้วยการสุ่ม โดยแบ่งออกเป็นข้อมูลฝึกสอน 70% และข้อมูลทดสอบ 30% เมื่อได้ข้อมูลฝึกสอน 70% แล้วผู้วิจัยได้ใช้วิธีการแบบ Cross-Validation Test ซึ่งเป็นวิธีที่ได้รับความนิยมสำหรับการแบ่งข้อมูลเพื่อวัดประสิทธิภาพของตัวแบบ เนื่องจากผลลัพธ์ที่ได้มีความน่าเชื่อถือ โดยหลักในการแบ่งข้อมูลด้วยวิธีนี้จะเริ่มจากการกำหนดค่า K หรือการแบ่งข้อมูลออกเป็น K ส่วนเท่า ๆ กัน ตัวอย่างเช่น กำหนดให้ K = 5 (5 Fold Cross-Validation) ข้อมูลจะถูกแบ่งออกเป็น 5 ส่วน โดยในแต่ละส่วนจะมีจำนวนข้อมูลเท่า ๆ กัน จากนั้นจะใช้ข้อมูล 4 ส่วนทำการเรียนรู้ 1 ส่วนใช้ทดสอบ



ภาพ 3 การแบ่งข้อมูลเพื่อวัดประสิทธิภาพของตัวแบบ

ประสิทธิภาพสลับกันไปจนครบทุกชุดข้อมูล ดังภาพ สามารถอธิบายลักษณะการแบ่งข้อมูลเพื่อทดสอบประสิทธิภาพของตัวแบบได้ดังตัวอย่างเช่น มีข้อมูลตัวอย่างทั้งหมดจำนวน 100 ตัวอย่าง ข้อมูลจะถูกแบ่งออกเป็น 5 ชุด โดยในแต่ละชุดจะมีข้อมูลทั้งหมด 20 ตัวอย่าง หลังจากนั้นทำการทดสอบประสิทธิภาพดังนี้

รอบที่ 1 นำข้อมูลชุดที่ 2,3,4,5 เรียนรู้เพื่อสร้างแบบจำลอง และใช้ข้อมูลชุดที่ 1 ทดสอบ

รอบที่ 2 นำข้อมูลชุดที่ 1,3,4,5 เรียนรู้เพื่อสร้างแบบจำลอง และใช้ข้อมูลชุดที่ 2 ทดสอบ

รอบที่ 3 นำข้อมูลชุดที่ 1,2,4,5 เรียนรู้เพื่อสร้างแบบจำลอง และใช้ข้อมูลชุดที่ 3 ทดสอบ

รอบที่ 4 นำข้อมูลชุดที่ 1,2,3,5 เรียนรู้เพื่อสร้างแบบจำลอง และใช้ข้อมูลชุดที่ 4 ทดสอบ

รอบที่ 5 นำข้อมูลชุดที่ 1,2,3,4 เรียนรู้เพื่อสร้างแบบจำลอง และใช้ข้อมูลชุดที่ 5 ทดสอบ

จากตัวอย่างจะได้ประสิทธิภาพของตัวแบบทั้งหมด 5 แบบจำลอง จากนั้นทำการหาค่าเฉลี่ยความถูกต้องของแบบจำลองทั้ง 5 ดังนั้นวิธีการนี้จึงมีความน่าเชื่อถือสูงสำหรับการวัดประสิทธิภาพของตัวแบบเนื่องจากข้อมูลทุกตัวจะถูกใช้ในการทดสอบประสิทธิภาพทำให้ไม่เกิดความเอนเอียงของข้อมูล

การวัดประสิทธิภาพตัวแบบ

การวัดประสิทธิภาพตัวแบบการจำแนกประเภทข้อมูล สามารถทำได้ด้วยกันหลายวิธี เช่น Confusion Matrix, Precision and Recall, F-Measure, Accuracy และ ROC Graph โดยในงานวิทยานิพนธ์นี้ได้เลือกใช้ค่าความถูกต้องมาเป็นตัววัดประสิทธิภาพของตัวแบบเนื่องจากทำการวัด

ความถูกต้องในการทำนาย โดยไม่สนใจว่าค่าที่ได้นั้นจะให้ค่าของลาเบลเป็นจริงหรือเป็นเท็จ เพียงแต่ตัวแบบสามารถทำนายได้ตรงกับลาเบล

ค่าความถูกต้อง (Accuracy) คือสัดส่วนความถูกต้องในการทำนายของทุก ๆ คลาส สามารถคำนวณหาได้ดังสมการ (9)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (9)$$

โดย

True Positive (TP) คือ สิ่งที่โปรแกรมทำนายว่า จริง และมีค่าเป็น จริง

True Negative (TN) คือ สิ่งที่โปรแกรมทำนายว่า ไม่จริง และมีค่าเป็น ไม่จริง

False Positive (FP) คือ สิ่งที่โปรแกรมทำนายว่า จริง แต่มีค่าเป็น ไม่จริง

False Negative (FN) คือ สิ่งที่โปรแกรมทำนายว่า ไม่จริง แต่มีค่าเป็น จริง

งานวิจัยที่เกี่ยวข้อง

ข้อมูลล็อกไฟล์นั้นสามารถนำไปสู่การตรวจจับการบุกรุกทางเครือข่ายได้ด้วยการนำล็อกไฟล์มาใช้ในการวิเคราะห์เพื่อหาความผิดปกตินั้นได้มีนักวิจัยได้ทำการวิจัยเกี่ยวกับล็อกไฟล์ เช่น Ertam & Kaya (2018), Brandao & Georgieva (2020), Ryciak et al. (2022) และ Wadekar et al. (2019) ได้มีการนำข้อมูลจากล็อกไฟล์มาใช้ในการวิเคราะห์เพื่อหาความผิดปกติของการทำงานของระบบคอมพิวเตอร์ที่อาจจะถูกผู้ไม่ประสงค์ดีเข้ามาในระบบการทำงานด้วยวิธีการการเรียนรู้ของเครื่อง การตรวจจับการบุกรุก ในรูปแบบปกติจะเป็นการตรวจจับโดยอาศัยกฎ (Rule-based Intrusion Detection Systems หรือ Rule-based IDS) จะมีการทำงานที่รวดเร็วและใช้ทรัพยากรในการประมวลผลที่ต่ำ ดังนั้น rule-based IDS จะเหมาะสมกับสภาพแวดล้อมที่มีเซสชันการเชื่อมต่อจำนวนมากต่อวินาทีที่ การจราจรทางเครือข่ายปริมาณมาก แต่วิธีการแบบ rule-based IDS ค่อนข้างจะขาดความยืดหยุ่นไม่สามารถมาตรวจจับการโจมตีที่รูปแบบใหม่ที่เกิดขึ้นอย่างรวดเร็ว (Tufan et al., 2021) จึงมีการสร้างระบบ IDS ด้วยเทคนิคการเรียนรู้ของเครื่อง สำหรับการเรียนรู้ของเครื่องได้ถูกใช้ในการสร้างตัวแบบการวิเคราะห์แบบอัตโนมัติหาความสัมพันธ์ที่ซ่อนอยู่ของข้อมูลเข้าและข้อมูลออก เป็นเทคนิคที่ในการวิเคราะห์ข้อมูลจัดเป็นหนึ่งสาขาของปัญญาประดิษฐ์ ซึ่งมีแนวคิดในการทำงานคือ การฝึกสอนระบบเพื่อให้ระบบทำการตัดสินใจและเรียนรู้ที่จะทำการระบุรูปแบบต่าง ๆ โดยพยายามให้มนุษย์มาเกี่ยวข้องในการทำงานให้น้อยที่สุด (Halimaa & Sundarakantham, 2019)

ถึงแม้การใช้เทคนิคด้านการเรียนรู้ของเครื่องจะมีผลการตรวจจับสิ่งผิดปกติได้อย่างมีประสิทธิภาพ แต่การเรียนรู้ของเครื่องมักจะประสบปัญหาที่ผลการทำงานมีค่าอัตราของผลบวกปลอม (False Positive) ที่สูงเนื่องมาจากค่าสหสัมพันธ์ระหว่างข้อมูลเครือข่ายหรือคุณลักษณะบางอย่างไม่ถูกนำมาใช้หรือถูกใช้น้อยมากในการทำงานการเรียนรู้ของเครื่อง (Li et al., 2017) แต่ก็มีงานวิจัยที่แนะนำให้ใช้การวิเคราะห์แบบ multivariate correlation สำหรับข้อมูลที่ได้จากการสกัดข้อมูลการจราจรในเครือข่าย เช่น งานของ (Yu et al., 2011) เสนอวิธีการจำแนกหาการโจมตีแบบ Distributed Denial of Service (DDoS) จากการเข้าใช้งานเครือข่ายจำนวนมากที่เพิ่มอย่างกะทันหัน จากการวิเคราะห์กระแสข้อมูลการจราจรทางเครือข่ายที่เข้ามาของสัมประสิทธิ์สหสัมพันธ์ของข้อมูลการจราจรที่น่าสงสัย ซึ่งวิธีในการตรวจจับโดยใช้คุณลักษณะดังกล่าวสามารถช่วยเพิ่มความความแม่นยำในการตรวจจับได้สูงขึ้น แต่ก็ยังมีจุดอ่อนต่อรูปแบบการโจมตีที่มีการเปลี่ยนแปลงรูปคุณลักษณะการโจมตีในลักษณะเชิงเส้น เช่น ในงานของ Li et al. (2017) ในงานวิจัยนี้จะเสนอการใช้การแปลงค่าคุณลักษณะให้อยู่ในรูปพหุนามเพื่อมาช่วยในการตรวจจับการบุกรุกในเครือข่าย Tan et al. (2014) แนะนำการสกัดค่าคุณลักษณะความสัมพันธ์จะสามารถเพิ่มประสิทธิภาพการทำงานถ้าประยุกต์ใช้การแปลงเป็นค่าพหุนามเข้ามาร่วมด้วย Resende & Drummond (2018) ได้ทำการสำรวจถึงงานวิจัยด้าน IDS โดยใช้เทคนิคการสุ่มป่าไม้ในช่วงเวลาสิบปีที่ผ่านมา นักวิจัยจำนวนมากได้เสนอวิธีในการทำ IDS ที่หลายชนิด เพื่อจัดการกับการเพิ่มจำนวนและความซับซ้อนของภัยคุกคามทางระบบคอมพิวเตอร์ ในบริบทนี้เทคนิคการสุ่มป่าไม้มีประสิทธิภาพอย่างเห็นได้ชัดและได้รับความนิยมในการใช้สำหรับ IDS ซึ่งเทคนิคการสุ่มป่าไม้มีความสามารถในการจำแนกประเภท การคัดเลือกคุณลักษณะ การประมาณค่าเมตริกซ์ สำหรับจุดเด่นของเทคนิคการสุ่มป่าไม้ในงานด้าน IDS เพราะเทคนิคการสุ่มป่าไม้มีจุดเด่นเมื่อเทียบกับเทคนิคด้านการเรียนรู้ของเครื่องอื่น ๆ ดังต่อไปนี้ 1) ใช้เวลาในการเรียนรู้ที่ต่ำ 2) ทนทานต่อชุดข้อมูลที่ไม่สมดุล 3) มีความสามารถในการคัดเลือกคุณลักษณะที่เหมาะสมอยู่ภายในตัวเอง รวมถึงมีเมตริกซ์ที่จัดลำดับคุณลักษณะที่สำคัญ 4) สามารถจัดการคุณลักษณะที่เป็นแบบกลุ่ม หรือแบบต่อเนื่อง จุดเด่นที่ได้กล่าวมานี้จะเห็นได้ชัด เมื่อมีการวัดประสิทธิภาพเทียบกับเทคนิคการเรียนรู้ของเครื่องอื่น ในงานด้าน IDS เช่นในงาน Ferriyan et al. (2017) เทียบ Random Forest, Bayesian Net, Naïve Bayes, K-Nearest Neighbor และ Decision tree C4.5 พบว่า การสุ่มป่าไม้มีประสิทธิภาพดีที่สุด ตัวแบบของ Random Forest นั้นจะเป็นการรวมกันของต้นไม้ตัดสินใจหลายต้น สามารถใช้ทำงานจำแนกประเภท หรือการถดถอย ผลการทำงานของการสุ่มป่าไม้งาน จำแนกประเภทจะอาศัยผลการโหวตเสียงส่วนใหญ่จากต้นไม้ตัดสินใจ ถ้าเป็นการถอถอยผลจะเป็นค่าเฉลี่ยจากต้นไม้ตัดสินใจ

เพื่อให้สามารถทำความเข้าใจในงานวิจัยที่เกี่ยวข้องได้อย่างสะดวกจึงขอสรุปเป็นตารางสรุปงานวิจัย ดังตาราง 1

ตาราง 1 แสดงสรุปงานวิจัยที่เกี่ยวข้อง

ลำดับ	ชื่องานวิจัย	จุดประสงค์	วิธีการวิจัย	ผลการวิจัย
1	Classification of Firewall Log Files with Multiclass Support Vector Machine (Ertam & Kaya, 2018)	การจัดประเภทไฟล์บันทึกของไฟร์วอลล์โดยใช้ตัวแยกประเภท multiclass support vector machine (SVM) และประเมินประสิทธิภาพของตัวแยกประเภทโดยใช้ฟังก์ชันการเปิดใช้งานที่แตกต่างกัน วัตถุประสงค์คือเพื่อให้ได้การตอบสนอง SVM ที่ดีที่สุดสำหรับค่าคะแนน F1 และสร้างเส้นโค้งลักษณะการทำงานของเครื่องรับ (ROC) สำหรับแต่ละคลาส	การจำแนกประเภทไฟล์บันทึกของไฟร์วอลล์โดยใช้ตัวแยกประเภท multiclass support vector machine (SVM) วิธีการเกี่ยวข้องกับการรับบันทึกผ่านไฟร์วอลล์ บันทึกที่ใช้นำมาจากอุปกรณ์ไฟร์วอลล์ Palo Alto 5020 ที่ใช้ในมหาวิทยาลัย Firat 65532 เร็กคอร์ดและบันทึกผลลัพธ์ประมาณ 30 วินาที มีการใช้วิธีการ SVM Linear, Polynomial, sigmoid และ Radial Basis Function (RBF) 4 วิธี และทำการวัดประสิทธิภาพด้วย F1 score, Precision, recall และ ROC	เสนอการใช้ตัวแยกประเภท multiclass support vector machine (SVM) สำหรับการจำแนกประเภทไฟล์บันทึกของไฟร์วอลล์เพื่อประเมินประสิทธิภาพของตัวแยกประเภท multiclass support vector machine (SVM) โดยใช้ฟังก์ชันที่แตกต่างกัน และวัดประสิทธิภาพโดยใช้ sensitivity, recall, and F1 Score เปรียบเทียบฟังก์ชันการใช้ SVM ที่ดีที่สุดสำหรับ F1 Score
2	Log Files Analysis for Network	วัตถุประสงค์ นำเสนอระบบตรวจจับการบุกรุกตามบันทึก	วิธีการวิจัยเกี่ยวข้องกับขั้นตอนต่อไปนี้	เสนอระบบตรวจจับการบุกรุกตามบันทึก (LIDS) ที่มีประสิทธิภาพเพื่อ

Intrusion Detection (Brandao & Georgieva, 2020)	(LIDS) ที่มีประสิทธิภาพเพื่อคาดการณ์ว่าบันทึกเครือข่ายเป็นการโจมตีหรือไม่ ระบบได้รับการออกแบบมาเพื่อตรวจสอบบันทึกจากแหล่งต่าง ๆ และระบุการโจมตีทางไซเบอร์โดยใช้เทคนิคการเรียนรู้ของเครื่อง ระบบที่เสนอนี้แสดงด้วยชุดข้อมูลไฟล์บันทึกที่มีป้ายกำกับที่ใหญ่ที่สุดที่เปิดเผยต่อสาธารณะ KDD Cup 1999	1. การรวมบันทึกจากแหล่งต่าง ๆ ไว้ในแดชบอร์ดเดียว 2. การกำหนดคุณลักษณะที่เลือกปฏิบัติมากที่สุดสำหรับการทำนายการโจมตี 3. ใช้วิธีการเลือกคุณลักษณะ 2 วิธี ได้แก่ การเลือกคุณลักษณะที่เหมาะสมที่สุด (OFS) และการวิเคราะห์ปัจจัย (FA) 4. เปรียบเทียบการทดสอบเทคนิคการเรียนรู้ของเครื่องบางอย่างสำหรับการทำนายการโจมตี โดยแผนผังการตัดสินใจเป็นผู้ชนะ 5. แสดงภาพระบบที่นำเสนอด้วยชุดข้อมูลไฟล์บันทึกที่มีป้ายกำกับที่ใหญ่ที่สุดที่เปิดเผยต่อสาธารณะ KDD Cup 1999 ข้อจำกัดของระบบตรวจจับการบุกรุกแบบดั้งเดิม (IDS) และความต้องการ	คาดการณ์ว่าบันทึกเครือข่ายเป็นการโจมตีหรือไม่ ระบบที่นำเสนอนี้แสดงด้วยชุดข้อมูล KDD Cup 1999 ผู้เขียนเปรียบเทียบการทดสอบเทคนิคการเรียนรู้ของเครื่องบางอย่างสำหรับการทำนายการโจมตี โดยแผนผังการตัดสินใจเป็นผู้ชนะ ผลลัพธ์แสดงว่าหากกำหนดคุณลักษณะที่เกี่ยวข้องมากที่สุด การเลือกลักษณนามจะมีความสำคัญน้อยกว่า การมุ่งเน้นไปที่การทำเหมืองข้อมูลมีความสำคัญมากขึ้นเพื่อให้ได้ผลลัพธ์ที่ดีที่สุด การกำจัดคุณลักษณะที่ไม่มีนัยสำคัญทำให้ปัญหาง่ายขึ้นและไม่กระทบต่ออัตราการตรวจจับ งานใน ขึ้นอยู่กับชุดข้อมูลโครงการ Honeynet ที่ไม่มีป้ายกำกับ หลังจากการทำให้ข้อมูลเป็นมาตรฐานผ่านโครงสร้าง

			IDS ใหม่ๆ ที่มุ่งเน้นไปที่เทคนิคการตรวจจับตามความผิดปกติและการใช้ในทางที่ผิด แนะนำว่าระบบตรวจจับการบุกรุกที่สมบูรณ์จะได้รับประโยชน์จากการประกอบด้วยแบบจำลองที่แตกต่างกันสำหรับงานที่แตกต่างกัน ซึ่งเมื่อรวมกันแล้วจะมีประสิทธิภาพมากขึ้นในการตรวจจับการโจมตีรูปแบบต่างๆ	Intrusion Detection Message Exchange Format (IDMEF) เชิงเส้นได้รับความแม่นยำ 98.2% โดยใช้ตัวจำแนกประเภทต้นไม้การตัดสินใจ
3	Combining Log Files and Monitoring Data to Detect Anomaly Patterns in a Data Center (Viola et al., 2022)	เสนอแบบจำลองที่รวมไฟล์บันทึกและข้อมูลการตรวจสอบเพื่อตรวจจับความผิดปกติใน data center เพื่อระบุความผิดปกติทั้งในระดับการบริการและพฤติกรรมระดับเครื่อง โดยเกี่ยวข้องกับการสร้างพจนานุกรมเฉพาะสำหรับล็อกไฟล์ โดยใช้เทคนิคต่าง ๆ เช่น wordclouds และการสร้าง	แนวคิดหลักคือการสร้างพจนานุกรมเฉพาะสำหรับไฟล์บันทึกแต่ละไฟล์ ซึ่งช่วยในการแยก n-grams ที่ผิดปกติในเมทริกซ์คุณลักษณะ มีการใช้เทคนิคหลายอย่างของ Natural Language Processing เช่น wordclouds และ Topic modeling เพื่อทำให้พจนานุกรมดังกล่าวสมบูรณ์ยิ่งขึ้น อัลกอริทึมการจัดกลุ่มถูกนำไปใช้กับเมทริกซ์คุณลักษณะเพื่อ	ผลการทดลองที่ดำเนินการใน data center จริงพร้อมกับไฟล์บันทึกและข้อมูลการตรวจสอบที่สามารถระบุลักษณะการทำงานของทรัพยากรจริงและเสมือนในการผลิต ตรวจสอบความสอดคล้องกันระหว่างความผิดปกติในไฟล์บันทึกและข้อมูลการตรวจสอบ เช่น การใช้หน่วยความจำลดลงหรือโหลดเครื่องเพิ่มขึ้น ผลลัพธ์มีแนวโน้มดีมาก และการรวมไฟล์

		แบบจำลองหัวข้อ เพื่อแยก n-grams ที่ผิดปกติ นอกจากนี้ยังใช้อัลกอริทึมการจัดกลุ่มกับเมทริกซ์คุณลักษณะ ในขณะที่ใช้การตรวจจับความผิดปกติของอนุกรมเวลาสำหรับข้อมูลเซนเซอร์	ระบุและจัดกลุ่มความผิดปกติประเภทต่าง ๆ ในอีกด้านหนึ่ง มีการใช้เทคนิคการตรวจจับความผิดปกติของอนุกรมเวลากับข้อมูลเซ็นเซอร์เพื่อรวมปัญหาที่พบในไฟล์บันทึกกับปัญหาที่เก็บไว้ในข้อมูลการตรวจสอบ ผลการทดลองที่ดำเนินการใน data center จริง พร้อมกับไฟล์บันทึกและข้อมูลการตรวจสอบที่สามารถระบุลักษณะการทำงานของทรัพยากรจริงและเสมือนในการผลิต	บันทึกและข้อมูลการตรวจสอบสามารถนำไปสู่ผลลัพธ์ที่สำคัญในการตรวจจับความผิดปกติใน data center อย่างไรก็ตาม ตัวแบบจำเป็นต้องผสานรวมความเชี่ยวชาญของผู้ดูแลไซต์เพื่อพิจารณาสถานการณ์ที่สำคัญทั้งหมดใน data center และทำความเข้าใจผลลัพธ์อย่างเหมาะสม
4	Feature analysis of encrypted malicious traffic (Shekhawat, 2019)	การวิเคราะห์คุณสมบัติต่าง ๆ ที่ใช้กันโดยทั่วไปเพื่อแยกแยะทราฟฟิกเครือข่ายที่เป็นอันตรายที่เข้ารหัสจากทราฟฟิกที่ไม่ปลอดภัย ในกรณีที่คีย์ถอดรหัสไม่สามารถใช้งานได้ โดยเฉพาะอย่างยิ่ง ใช้การเรียนรู้ของเครื่องเพื่อวิเคราะห์คุณสมบัติการรับส่ง	1. วิเคราะห์คุณสมบัติต่างๆ ที่ใช้กันโดยทั่วไปเพื่อแยกทราฟฟิกเครือข่ายที่เป็นอันตรายที่เข้ารหัสออกจากทราฟฟิกที่ไม่ร้ายแรงที่เข้ารหัส 2. ใช้อัลกอริธึมการเรียนรู้ของเครื่องสามแบบ ได้แก่ เครื่องสนับสนุนเวกเตอร์ (SVM) ป่าสุ่ม (RF) และการเพิ่มการไล่ระดับสีมาก (XG-Boost)	การทดลองแสดงให้เห็นว่าวิธีการที่นำเสนอมีความแม่นยำสูงสุดเกือบ 99% ซึ่งดีกว่า SVM เชิงเส้นอย่างมาก และมีความแม่นยำเกิน 98.5% ด้วยคุณลักษณะเพียง 6 ประการ โดยรวมแล้ว ผลลัพธ์ของบทความนี้แสดงให้เห็นถึงประสิทธิภาพของการใช้เทคนิคการเรียนรู้ของเครื่องเพื่อวิเคราะห์

		ข้อมูลเครือข่ายที่เข้ารหัส วัตถุประสงค์คือเพื่อแสดงให้เห็นว่าสามารถรับข้อมูลเกี่ยวกับคุณลักษณะได้โดยตรงจากตัวแบบการเรียนรู้ของเครื่องเอง ซึ่งมีความน่าเชื่อถือมากกว่าการพึ่งพาผู้เชี่ยวชาญของมนุษย์ในการพิจารณาคุณลักษณะที่เป็นประโยชน์และให้ข้อมูลมากที่สุดในโดเมนปัญหานี้	เพื่อวิเคราะห์คุณสมบัติการรับส่งข้อมูลเครือข่ายที่เข้ารหัส 3. ใช้เมตริกการประเมินเพื่อวัดผลที่ได้รับจากการทดลอง แสดงให้เห็นว่าข้อมูลเกี่ยวกับคุณลักษณะสามารถรับได้โดยตรงจากตัวแบบการเรียนรู้ของเครื่องเอง ซึ่งมีความน่าเชื่อถือมากกว่าการพึ่งพาผู้เชี่ยวชาญของมนุษย์ในการพิจารณาคุณลักษณะที่มีประโยชน์และให้ข้อมูลมากที่สุดในโดเมนปัญหานี้ โดยรวมแล้ว วิธีการวิจัยเกี่ยวข้องกับการใช้เทคนิคการเรียนรู้ของเครื่องเพื่อวิเคราะห์คุณลักษณะการรับส่งข้อมูลเครือข่ายที่เข้ารหัส และประเมินผลโดยใช้เมตริกที่เหมาะสม	คุณสมบัติการรับส่งข้อมูลเครือข่ายที่เข้ารหัสเพื่อตรวจหาการรับส่งข้อมูลที่เป็นอันตราย
5	Benchmark for filter methods for feature	เพื่อเปรียบเทียบประสิทธิภาพของวิธีการกรอง โดยคำนึงถึงทั้งเวลาทำงานและความแม่นยำใน	การวิเคราะห์และเปรียบเทียบประสิทธิภาพของวิธีการกรอง สำหรับการคัดเลือกคุณสมบัติในการเรียนรู้	ไม่มีกลุ่มวิธีการกรองที่มีประสิทธิภาพดีกว่าวิธีอื่นๆ เสมอไป แนะนำเกี่ยวกับวิธีการกรองที่ทำงานได้ดีกับ

	selection in high-dimensional classification data (Bommert, 2020)	การทำนายเมื่อรวมกับวิธีการจำแนกประเภท การวิเคราะห์ขึ้นอยู่กับชุดข้อมูลการจำแนกมิติสูง 16 ชุด	ของเครื่อง การวิเคราะห์ขึ้นอยู่กับชุดข้อมูลการจำแนกมิติสูง 16 ชุด และผู้เขียนใช้แพ็คเกจการเรียนรู้ของเครื่อง R เพื่อดำเนินการวิเคราะห์ นำเสนอวิธีการกรองสองประเภท ได้แก่ วิธีการคำนวณคะแนนสำหรับคุณสมบัติทั้งหมด จากนั้นเลือกคุณสมบัติที่มีคะแนนสูงสุด และวิธีที่เลือกคุณสมบัติซ้ำ ๆ	ชุดข้อมูลจำนวนมาก ยังระบุกลุ่มของตัวกรองที่คล้ายคลึงกันตามลำดับที่ตัวกรองเหล่านี้จัดลำดับคุณลักษณะ
6	Machine Learning based Intrusion Detection System (Halimaa & Sundarakantham, 2019)	เพื่อเปรียบเทียบประสิทธิภาพของ ML 2 เทคนิค ในการตรวจจับการบุกรุก	ทดลองเปรียบเทียบประสิทธิภาพของ SVM และ Naïve Bayes กับชุดข้อมูล NSL-KDD จำนวน 19,000 ตัวอย่าง	ผลการวัดประสิทธิภาพ SVM มีความแม่นยำ มากกว่า Naïve Bayes IDS + ML
7	งานวิทยานิพนธ์นี้	พัฒนาเฟรมเวิร์คสำหรับการค้นหาการคัดเลือกคุณลักษณะ		ไม่ได้เปรียบเทียบประสิทธิภาพความถูกต้องกับวิธีการจำแนกประเภทแบบ

	ของข้อมูลและค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมที่สุดของชุดข้อมูลล็อกไฟล์ด้วยการทำงานของวิธีการการสร้างคุณลักษณะพหุนามนำมาร่วมกับขั้นตอนวิธีเชิงพันธุกรรมเพื่อลดขนาดและมิติของข้อมูลในการจัดเก็บล็อกไฟล์	อื่น เนื่องจากเป็นการพัฒนาเฟรมเวิร์คสำหรับการค้นหาการคัดเลือกคุณลักษณะของข้อมูลและค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมที่สุด

บทที่ 3

วิธีดำเนินการวิจัย

สำหรับล็อกไฟล์ ที่เกิดจากการเก็บข้อมูลกิจกรรมที่เกิดขึ้นภายในเครือข่ายขององค์กร ที่จำเป็นต้องมีการเก็บข้อมูลตามที่กำหนดด้านกฎหมายการกระทำความผิดที่เกี่ยวกับคอมพิวเตอร์ ซึ่งจะใช้สืบค้นการบุกรุกที่อาจเกิดขึ้น สำหรับระบบตรวจสอบการตรวจจับการบุกรุกทางเครือข่าย ซึ่งจะใช้การวิเคราะห์ข้อมูลจากล็อกไฟล์ และ ไม่อาจทราบได้ว่าคุณลักษณะ ใดมีความสำคัญต่อการวิเคราะห์ของการบุกรุกทางเครือข่าย ในด้านการศึกษาการเลือกคุณลักษณะมีหลากหลายวิธี แต่ในบางครั้งไม่สามารถดึงความสัมพันธ์ที่ซ่อนอยู่ระหว่างคุณลักษณะได้ จึงมีการประยุกต์ใช้เทคนิคการสร้างคุณลักษณะพหุนาม ในการทำการบุกรุกทางเครือข่าย และ พบว่าช่วยเพิ่มประสิทธิภาพการตรวจจับการบุกรุกทางเครือข่าย

ซึ่งในการวิจัยนี้เสนอการปรับปรุงวิศวกรรมคุณลักษณะสำหรับตัวจำแนกประเภทแบบต้นไม้สำหรับการตรวจจับการบุกรุกทางเครือข่าย ได้เสนอเฟรมเวิร์คด้านการบุกรุกทางเครือข่ายที่ประยุกต์ใช้เทคนิคการเลือกคุณลักษณะ และ เทคนิคพหุนาม โดยในบทที่ 3 จะมีหัวข้อที่เกี่ยวข้องดังนี้

3.1 ชุดข้อมูลที่ใช้ในการวิจัย

3.1.1 ชุดข้อมูลการบุกรุกทางเครือข่าย

3.1.2 ชุดข้อมูลที่ไม่เกี่ยวข้องกับการบุกรุกทางเครือข่าย

3.2 เฟรมเวิร์คและขั้นตอนวิธีที่นำเสนอในการวิจัย

3.2.1 เฟรมเวิร์คการคัดเลือกคุณลักษณะข้อมูลที่ใช้เทคนิคทางสถิติกับเทคนิคพหุนาม

3.2.2 ขั้นตอนวิธีการคัดเลือกคุณลักษณะข้อมูลที่ใช้เทคนิคพหุนามร่วมกับขั้นตอนวิธีเชิงพันธุกรรม

3.2.3 การประยุกต์ใช้ขนาดของล็อกไฟล์เพื่อการเลือกไฮเปอร์พารามิเตอร์และแบบจำลองที่เหมาะสมที่สุด

3.1 ชุดข้อมูลที่ใช้ในการวิจัย

ในการวิจัยนี้ใช้งานชุดข้อมูลทั้งหมด 4 ชุดข้อมูล โดยใช้ชุดข้อมูลสาธารณะที่เผยแพร่สำหรับงานวิจัยด้านการจำแนก แบบมีการป้ายกำกับ แบ่งเป็น ชุดข้อมูลแรกคือ ชุดข้อมูลด้านการบุกรุกเครือข่ายจำนวน 2 ชุด ประกอบด้วย ชุดข้อมูลแรกคือชุดข้อมูลไอโอที และชุดข้อมูลที่สองคือ

ชุดข้อมูลเอ็นเอสแอลเคดีดี และ ชุดข้อมูลจากไซคิทเลิร์น จำนวน 2 ชุดประกอบด้วย ชุดข้อมูลมะเร็ง และชุดข้อมูลที่สองลายมือ ชุดข้อมูลทั้ง 4 ชุดข้อมูลเป็นชุดข้อมูลสาธารณะ ที่งานวิจัยส่วนใหญ่เลือกใช้เป็นข้อมูลสำหรับวัดประสิทธิภาพในการทำงานของขั้นตอนกระบวนการของแต่ละงานวิจัยประกอบด้วยชุดข้อมูล ดังนี้

ตาราง 2 แสดงชุดข้อมูลที่ใช้ในงานวิจัย

ชุดข้อมูล	จำนวนคุณลักษณะ	จำนวนแถวที่บันทึก	จำนวนทาร์เก็ต	ค่าที่ขาดหายไป
ไอโอที	83	625,785	2	Yes
เอ็นเอสแอลเคดีดี	41	47,735	2	Yes
มะเร็ง	30	569	2	No
ลายมือ	64	1,797	10	No

3.1.1 ชุดข้อมูลการบุกรุกทางเครือข่าย

1) ชุดข้อมูลไอโอที คือ ชุดข้อมูลการบุกรุกเครือข่าย Internet of Things (IoT) เป็นชุดข้อมูลการรับส่งข้อมูลเครือข่ายที่ออกแบบมาโดยเฉพาะสำหรับการศึกษาและประเมินเทคนิคการตรวจจับการบุกรุกในสภาพแวดล้อม IoT ด้วยจำนวนอุปกรณ์ IoT ที่เพิ่มจำนวนขึ้นอย่างรวดเร็ว การรับรองความปลอดภัยของอุปกรณ์จึงกลายเป็นปัญหาเร่งด่วน ระบบตรวจจับการบุกรุกเครือข่ายทั่วไปอาจไม่เหมาะที่จะจัดการกับลักษณะเฉพาะของเครือข่าย IoT เช่น ทรัพยากรการประมวลผลที่จำกัดและโปรโตคอลการสื่อสารที่หลายชนิด ด้วยเหตุนี้ จึงมีความต้องการชุดข้อมูลพิเศษเพิ่มขึ้น ซึ่งสามารถใช้ในการพัฒนาและประเมินเทคนิคการตรวจจับการบุกรุกเฉพาะของ IoT

ชุดข้อมูลการบุกรุกเครือข่าย IoT มีจุดมุ่งหมายเพื่อรวบรวมข้อมูลการรับส่งข้อมูลเครือข่ายที่ครอบคลุม และเป็นปัจจุบันที่มาจากอุปกรณ์ IoT ชุดข้อมูลนี้ประกอบด้วยทั้งการจราจรทางคอมพิวเตอร์แบบปกติและแบบมีการโจมตีประเภทต่างๆ เช่น Distributed Denial of Service (DDoS) การสแกน และการติดมัลแวร์ สามารถใช้ชุดข้อมูลเพื่อฝึกสอนและประเมินตัวแบบแมชชีนเลิร์นนิงที่ออกแบบมาโดยเฉพาะสำหรับการตรวจจับภัยคุกคามด้านความปลอดภัยในสภาพแวดล้อม IoT มีข้อมูลประเภท ดังนี้

-การรับส่งข้อมูลจากอุปกรณ์ IoT ต่างๆ ชุดข้อมูลอาจมีการรับส่งข้อมูลเครือข่ายที่สร้างโดยอุปกรณ์ IoT เช่น เครื่องใช้ในบ้านอัจฉริยะ อุปกรณ์สวมใส่ และเซ็นเซอร์ ของเครือข่าย IoT

-โปรโตคอลการสื่อสาร อุปกรณ์ IoT มักจะใช้โปรโตคอลการสื่อสาร เช่น MQTT, CoAP และ HTTP รวมถึงการรับส่งข้อมูลจากโปรโตคอลต่างๆ เหล่านี้ เพื่อช่วยให้นักวิจัยพัฒนาเทคนิคการตรวจจับการบุกรุกที่ปรับให้เหมาะกับความต้องการเฉพาะของเครือข่าย IoT

-ชุดข้อมูลรวบรวมภายใต้เงื่อนไขของเครือข่ายที่ใช้งานจริง เพื่อให้ระบบตรวจจับการบุกรุกต้องเผชิญในสภาพแวดล้อม IoT มีความแม่นยำมากขึ้น

-ข้อมูลที่มีป้ายกำกับ การเชื่อมต่อเครือข่าย แต่ละรายการในชุดข้อมูลจะมีป้ายกำกับว่าปกติหรือเป็นการโจมตี โดยประเภทหลังจะถูกจัดประเภทเพิ่มเติมตามประเภทการโจมตีต่างๆ ป้ายกำกับนี้ช่วยให้สามารถฝึกสอนตัวแบบแมชชีนเลิร์นนิงภายใต้การดูแลและประเมินประสิทธิภาพได้

-สถานการณ์การโจมตีที่หลายประเภท ชุดข้อมูลรวมถึงการโจมตีประเภทต่างๆ เช่น DDoS การสแกน และการติดมัลแวร์ เพื่อให้ครอบคลุมภัยคุกคามความปลอดภัยที่หลายประเภทในเครือข่าย IoT (รายละเอียดของชุดข้อมูลไอโอที อยู่ในภาคผนวก)

ตาราง 3 ลักษณะรูปแบบของข้อมูลที่มีการบันทึกในชุดข้อมูลไอโอที

sbytes	dbytes	rate	....	sinpkt	dinpkt	sjit	attack
530	268	10.45904	....	151.2431	270.8868	10042.87	Normal
530	354	8.834441	....	213.7388	261.252	13125.11	Normal
7954	354	7.570899	....	213.3676	381.2934	18158.23	Normal
2516	354	11.60377	....	157.7034	195.4947	8764.724	Normal
816	1172	17.27864	....	109.3193	124.9329	5929.212	Normal
534	268	21.00305	....	79.35356	120.1914	4013.392	Normal
826	1266	12.3758	....	170.4819	159.0706	11933.07	Normal
92	0	0.016668	....	59998.2	0	0	Normal
534	268	19.80734	....	84.02333	123.7992	4064.837	Normal
534	354	22.38887	....	84.24267	98.50171	4716.887	Normal

2) ชุดข้อมูลเอ็นเอสแอลเคดีดี เป็นชุดข้อมูลสาธารณะที่ใช้กันอย่างแพร่หลายสำหรับการวิจัยการตรวจจับการบุกรุกเครือข่าย เป็นชุดข้อมูลที่ถูกปรับปรุงจากชุดข้อมูล KDD Cup '99 รุ่นก่อนหน้า ซึ่งสร้างขึ้นสำหรับการแข่งขันระดับนานาชาติ ชุดข้อมูล KDD Cup '99 เดิมมีปัญหาหลายอย่าง เช่น บันทึกซ้ำซ้อนและการกระจายการโจมตีที่ไม่สมจริง ซึ่งทำให้ยากต่อการใช้งานสำหรับการวิจัยการตรวจจับการบุกรุก ชุดข้อมูลเอ็นเอสแอลเคดีดี ได้รับการพัฒนาเพื่อแก้ไขปัญหาเหล่านี้และจัดเตรียมชุดข้อมูลที่สะอาดและสมดุลมากขึ้นสำหรับนักวิจัย ชุดข้อมูล เอ็นเอสแอลเคดีดี ประกอบด้วยข้อมูลเกี่ยวกับการเชื่อมต่อเครือข่ายที่รวบรวมจากสภาพแวดล้อมเครือข่ายทางทหาร

จำลอง แต่ละการเชื่อมต่อหรืออินสแตนซ์จะแสดงด้วยชุดคุณลักษณะ 41 รายการ ซึ่งรวมถึงแอตทริบิวต์การรับส่งข้อมูลเครือข่ายต่างๆ เช่น ประเภทโปรโตคอล บริการ ระยะเวลา และจำนวนไบต์ที่ส่ง นอกจากนี้ การเชื่อมต่อแต่ละรายการจะมีป้ายกำกับว่า 'ปกติ' หรือ 'โจมตี' โดยประเภทหลังแบ่งประเภทเพิ่มเติมออกเป็นสี่ประเภทหลัก การปฏิเสธการให้บริการ (DoS), ระยะไกลไปยังท้องถิ่น (R2L), ผู้ใช้ถึงรูท (U2R) และ การตรวจสอบ ประเภทการโจมตีเหล่านี้ครอบคลุมถึงการบุกรุกเฉพาะต่างๆ ที่สามารถตรวจจับได้โดยระบบตรวจจับการบุกรุกเครือข่าย (NIDS) คุณสมบัติหลักของชุดข้อมูล เอ็นเอสแอลเคดีดี สามารถแบ่งออกได้เป็น

คุณลักษณะพื้นฐาน คุณลักษณะเหล่านี้ได้รับมาจากแพ็กเก็ตเครือข่ายดิบ เช่น จำนวนไบต์ที่ส่ง ประเภทโปรโตคอล (TCP, UDP หรือ ICMP) และบริการ (เช่น http, ftp หรือ telnet)

คุณสมบัติเนื้อหาคุณสมบัติเหล่านี้ดึงมาจากส่วนข้อมูลของแพ็กเก็ตเครือข่าย เช่น จำนวนครั้งของการพยายามเข้าสู่ระบบที่ล้มเหลว หรือจำนวนของแพ็กเก็ตเร่งด่วน

คุณลักษณะการรับส่งข้อมูล คุณลักษณะเหล่านี้ได้รับมาจากการรับส่งข้อมูลเครือข่ายในช่วงเวลาที่กำหนด รวมถึงคุณลักษณะต่างๆ เช่น จำนวนการเชื่อมต่อกับโฮสต์เดียวกันหรือบริการเดียวกัน (รายละเอียดของชุดข้อมูลเอ็นเอสแอลเคดีดี อยู่ในภาคผนวก)

ตาราง 4 ลักษณะรูปแบบของข้อมูลที่มีการบันทึกในชุดข้อมูลเอ็นเอสแอลเคดีดี

protocol_type	service	flag	src_bytes	dst_bytes	srv_count	class
tcp	ftp_data	SF	491	0	2	normal
udp	other	SF	146	0	1	normal
tcp	private	S0	0	0	6	neptune
tcp	http	SF	232	8153	5	normal
tcp	http	SF	199	420	32	normal
tcp	private	REJ	0	0	19	neptune
tcp	private	S0	0	0	9	neptune
tcp	private	S0	0	0	16	neptune
tcp	private	S0	0	0	8	neptune
tcp	private	REJ	0	0	12	neptune

นอกเหนือจากคุณลักษณะ 41 รายการเหล่านี้แล้ว ชุดข้อมูลเอ็นเอสแอลเคดีดี ยังมีป้ายกำกับคลาสที่ระบุว่าการจราจรในเครือข่ายปกติหรือโดนโจมตี ในการโจมตีแบ่งออกเป็นสี่กลุ่มหลัก

เพิ่มเติม ได้แก่ Denial of Service (DoS), Probing, User to Root (U2R) และ Remote to Local (R2L)

3.1.2 ชุดข้อมูลที่ไม่เกี่ยวข้องกับการบุกรุกทางเครือข่าย

1). ชุดข้อมูลมะเร็ง เป็นชุดข้อมูลที่ใช้กันอย่างแพร่หลายสำหรับการศึกษาการจัดประเภทของมะเร็งเต้านม สร้างขึ้นโดย Dr. William H. Wolberg และเพื่อนร่วมงานที่มหาวิทยาลัยวิสคอนซิน ชุดข้อมูลประกอบด้วย 569 ตัวอย่าง โดยแต่ละรายการมีคุณลักษณะ 32 รายการ ประกอบด้วย หมายเลขประจำตัว การวินิจฉัย (M = มะเร็ง, B = ไม่เป็น) และคุณลักษณะอีก 30 คุณลักษณะ ที่คำนวณจากภาพดิจิทัลของการตรวจสอบทางการแพทย์ที่ชี้บ่งสาเหตุของก้อนผิดปกติในร่างกาย โดยการใช้เข็มเจาะดูดเอาชิ้นเนื้อออกมา FNA (Fine Needle Aspiration) คุณลักษณะแบ่งออกเป็นสามกลุ่ม ค่าเฉลี่ย ข้อผิดพลาดมาตรฐาน และแย่ที่สุด แต่ละกลุ่มประกอบด้วยคุณสมบัติ 10 ประการ

ตาราง 5 ลักษณะข้อมูลที่อยู่ในชุดข้อมูลมะเร็ง

ID	Radius	Texture	Perimeter	Area	Smoothness	Compactness	Concavity	Class (M-malignant B-benign)
1	28.1	10.5	122.8	1001	0.1184	0.2778	0.3	M
2	27	10.4	120.1	811	0.1184	0.2778	0.2867	M
3	25.4	10.4	104.1	569	0.1184	0.2778	0.2778	M
4	22.7	10.3	85	367	0.1184	0.2778	0.2969	B
5	21.2	10.2	79.8	340	0.1184	0.2778	0.3	B

2) ชุดข้อมูลลายมือ มีอยู่ในไลบรารี ไซคิทเลิร์น ใน Python เป็นชุดข้อมูลที่ได้รับความนิยมสำหรับงานการเรียนรู้ของเครื่อง โดยเฉพาะอย่างยิ่งสำหรับ ฝึกสอนเทคนิคการจำแนกประเภทรูปภาพ ชุดข้อมูลประกอบด้วยรูปภาพระดับสีเทา 1,797 ภาพ ที่เขียนด้วยลายมือ แต่ละภาพแทนเลขหลักเดียวตั้งแต่ 0 ถึง 9 ความละเอียดของภาพ ภาพในชุดข้อมูลมีความละเอียด 8x8 พิกเซล ซึ่งทำให้มีขนาดค่อนข้างเล็กและประมวลผลได้ง่ายกว่าเมื่อเทียบกับภาพความละเอียดสูง ค่าความเข้มของพิกเซลในภาพอยู่ในช่วงตั้งแต่ 0 ถึง 16 โดยที่ 0 หมายถึงสีขาวและ 16 หมายถึงสีดำ ค่าเหล่านี้จะถูกเก็บเป็นจำนวนเต็ม แต่ละภาพในชุดข้อมูลมีป้ายกำกับที่สอดคล้องกัน ซึ่งเป็นตัวเลขจริง (0-9) ที่รูปภาพนั้นใช้แทน ป้ายกำกับเหล่านี้ใช้สำหรับงานการเรียนรู้ภายใต้การดูแล เช่น การจัดหมวดหมู่ รูปภาพจะถูกจัดเก็บเป็นอาร์เรย์ NumPy ทำให้ง่ายต่อการประมวลผลและจัดการข้อมูลโดยใช้

ไลบรารี Python เช่น NumPy, ไซคิทเลิร์น และ TensorFlow ชุดข้อมูลตัวเลขที่เขียนด้วยลายมือพร้อมใช้งานผ่านไลบรารี ไซคิทเลิร์น ทำให้ง่ายต่อการโหลดและใช้สำหรับการฝึกสอนและทดสอบตัวแบบแมชชีนเลิร์นนิง ชุดข้อมูลได้รับการประมวลผลล่วงหน้า รูปภาพได้รับการปรับขนาด และค่าความเข้มของพิกเซลได้รับการทำให้เป็นมาตรฐานในช่วงที่สอดคล้องกัน ขั้นตอนการประมวลผลล่วงหน้านี้ช่วยประหยัดเวลาและความพยายามเมื่อทำงานกับชุดข้อมูล ชุดข้อมูลตัวเลขที่เขียนด้วยลายมือเหมาะสำหรับงานด้านแมชชีนเลิร์นนิงและคอมพิวเตอร์วิทัศน์ต่างๆ รวมถึงการรู้จำตัวเลข การจัดหมวดหมู่รูปภาพ และเป็นขั้นตอนการประมวลผลล่วงหน้าสำหรับงานที่ซับซ้อนมากขึ้น เช่น การรู้จำอักขระด้วยแสง (OCR)

ตาราง 6 ลักษณะข้อมูลของชุดข้อมูลลายมือ

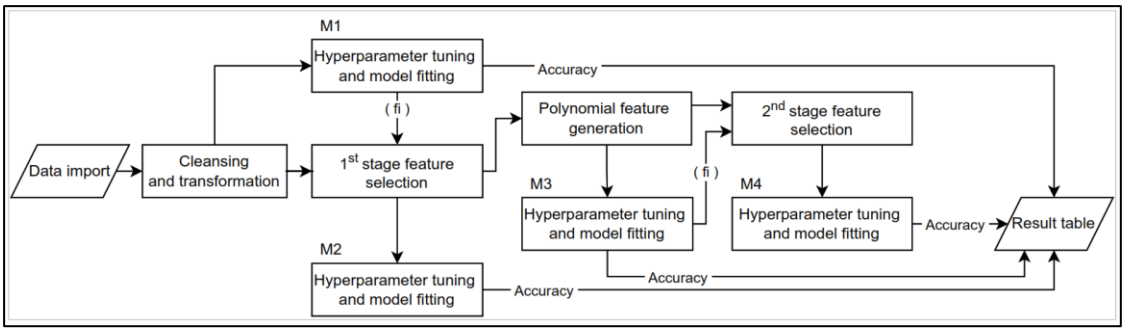
แถว	Column1	Column2	Column3	Column4	....	Column62	Column63	Column64	Class
1	0	1	6	15	....	1	0	0	0
2	0	0	10	16	....	3	0	0	0
3	0	0	8	15	....	0	0	0	7
4	0	0	0	3	....	2	0	0	4
5	0	0	5	14	....	7	0	0	6

ตารางนี้นำเสนอชุดข้อมูล และแสดงเฉพาะค่าความเข้มพิกเซล (คอลัมน์) สำหรับแต่ละภาพ มีค่าความเข้ม 64 พิกเซล (ตาราง 8x8) สำหรับแต่ละภาพในชุดข้อมูล ตารางนี้แสดงโครงสร้างของชุดข้อมูล โดยมีแถวแสดงรูปภาพ คอลัมน์แสดงความเข้มของพิกเซล และคอลัมน์ป้ายกำกับที่มีตัวเลขจริงที่ปรากฏในแต่ละรูปภาพ ในชุดข้อมูล คุณลักษณะแต่ละอย่างแสดงถึงค่าความเข้มของพิกเซลระดับสีเทาตั้งแต่ 0 ถึง 16 ค่าเป็นตัวเลข

3.2 เฟรมเวิร์คที่นำเสนอในการวิจัย

ในการวิจัยนี้มีการเสนอเฟรมเวิร์คและขั้นตอนวิธี 2 ประเภท คือ แบบที่ใช้เทคนิคทางสถิติกับเทคนิคพหุนาม และ แบบที่ใช้เทคนิคพหุนามร่วมกับวิธีเชิงพันธุกรรม

3.2.1 เฟรมเวิร์คการคัดเลือกคุณลักษณะข้อมูลที่ใช้เทคนิคทางสถิติกับเทคนิคพหุนาม จะมีขั้นตอนการทำงานดังแสดงในภาพ aa มีขั้นตอนการทำงานทั้งหมด 10 ขั้นตอนดังแสดงในภาพ



ภาพ 4 แสดงเฟรมเวิร์คการคัดเลือกคุณลักษณะข้อมูลที่ใช้เทคนิคทางสถิติกับเทคนิคพหุนาม

ขั้นตอนการทำงานสำหรับเฟรมเวิร์คการคัดเลือกคุณลักษณะข้อมูลที่ใช้เทคนิคทางสถิติกับเทคนิคพหุนาม

1) การนำข้อมูลชุดเข้า

การวิจัยที่มีการเปรียบเทียบประสิทธิภาพในการทำงานของผลลัพธ์ จะต้องมีการแบ่งข้อมูลเป็นข้อมูลฝึกสอน และข้อมูลทดสอบ ในอัตราส่วน 70:30

สำหรับขั้นตอนการนำเข้าข้อมูลที่เกี่ยวข้องที่จำเป็นสำหรับการวิจัยนี้ โดยข้อมูลที่นำเข้าเป็นข้อมูลที่ยังไม่ผ่านกระบวนการทำความสะอาดสำหรับข้อมูลที่มีความผิดปกติ เช่น ปัญหาข้อมูลสูญหายในบางแถว ดังภาพ 7 ข้อมูลแถวที่ 5 คอลัมน์ D ไม่มีข้อมูล หรือ ค่าว่าง ซึ่งข้อมูลแถวที่มีค่าว่างจะส่งผลต่อการทำงานของทดลองต้องดำเนินการแก้ปัญหานี้

ตาราง 7 แสดงรายละเอียดข้อมูล

ลำดับ	duration	Protocol_type	service	flag	Src_bytes	Dst_bytes	land
1	0	Tcp	ftp_data	SF	491	0	0
2	0	Udp	Other	SF	146	0	0
3	0	Tcp	Private	S0	0	0	0
4	0	Tcp	http		232	8153	0
5	0	Tcp	http	SF	199	420	0
6	0	Tcp	http	SF	0	0	0
7	0	Tcp	http	REJ	0	0	0
8	0	Tcp	Private	SF	0	0	0
9	0	Tcp	Private	SF	0	0	0
10	0	icmp	http	S0	0	0	0

ตาราง 7 แสดงตัวอย่างข้อมูลจากชุดข้อมูลชุดข้อมูลเอ็นเอสแอลเคดีดี เป็นข้อมูลการบุกรุกทางเครือข่ายจากล็อกไฟล์ เพียงบางส่วน เพื่อแสดงให้เห็นข้อมูลก่อนดำเนินการตรวจหาข้อมูลผิดปกติ

2) การทำความสะอาดข้อมูลและแปลงข้อมูล (Cleansing and transformation)

การทำความสะอาดข้อมูลและแปลงค่าข้อมูลเกี่ยวข้องกับการเตรียมชุดข้อมูล จะดำเนินการหาข้อมูลแถว หรือคอลัมน์ใดที่ไม่สมบูรณ์ เช่น มีค่าว่าง ค่าผิดปกติที่มีลักษณะความไม่สอดคล้องของข้อมูลในคอลัมน์เดียวกัน ดังภาพ 8 ที่มีการดำเนินการนำข้อมูลแถวที่มีค่าว่างออกจากชุดข้อมูล และการแปลงค่าให้อยู่ในรูปแบบที่เหมาะสม เช่น ข้อมูลที่มีค่าไม่ใช่ตัวเลข ดังภาพ เช่น คอลัมน์ D ค่าคอลัมน์ flag ที่มีค่าไม่ใช่ตัวเลขแต่เป็นตัวหนังสือประกอบด้วย SF,S0,REJ,S0 ต้องมีการแปลงรูปแบบข้อมูลเป็นตัวเลขดังแสดงภาพ 9 ที่จะมีคอลัมน์ protocol _type และ flag เป็นข้อมูลที่ไม่ใช่ตัวเลขจำเป็นต้องมีการแปลงข้อมูลดังแสดงในภาพ 10 ที่มีการเพิ่มคอลัมน์ protocol_type_tcp, protocol_type_udp, protocol_type_icmp

ตาราง 8 แสดงรายละเอียดข้อมูล

ลำดับ	duration	Protocol_type	service	flag	Src_bytes	Dst_bytes	land
1	0	Tcp	ftp_data	SF	491	0	0
2	0	Udp	Other	SF	146	0	0
3	0	Tcp	Private	S0	0	0	0
5	0	Tcp	http	SF	199	420	0
6	0	Tcp	http	SF	0	0	0
7	0	Tcp	http	REJ	0	0	0
8	0	Tcp	Private	SF	0	0	0
9	0	Tcp	Private	SF	0	0	0
10	0	icmp	http	S0	0	0	0

ตาราง 9 แสดงรายละเอียดข้อมูล

ลำดับ	duration	Protocol_type_tcp	Protocol_type_udp	Protocol_typeicmp	service	flag	Src_bytes
1	0	1	0	0	ftp_data	SF	491
2	0	0	1	0	Other	SF	146
3	0	1	0	0	Private	S0	0

ลำดับ	duration	Protocol_type_tcp	Protocol_type_udp	Protocol_typeicmp	service	flag	Src_bytes
5	0	1	0	0	http	SF	199
6	0	1	0	0	http	SF	0
7	0	1	0	0	http	REJ	0
8	0	1	0	0	Private	SF	0
9	0	1	0	0	Private	SF	0
10	0	0	0	1	http	S0	0

เมื่อชุดข้อมูลเอ็นเอสแอลเคดีดี ซึ่งมีคุณลักษณะที่เป็นหมวดหมู่อยู่ 3 คอลัมน์ ประกอบไปด้วย protocol_type, service และ flag เมื่อผ่านกระบวนการทำแปลงข้อมูลจำนวนคุณลักษณะจะถูกขยายเป็น 118 คุณลักษณะ ตามคลาสที่มีในคุณลักษณะที่เป็นหมวดหมู่

ตาราง 10 แสดงจำนวนต่าง ๆ ของชุดข้อมูลไอโอที

ประเภท		จำนวนแถว	จำนวนคุณลักษณะ	จำนวนในแต่ละประเภท	
				ปกติ	ผิดปกติ
ข้อมูลดิบ		20,000	86		
ข้อมูลสะอาด		20,000	77		
ข้อมูลแปลง	ข้อมูลฝึกสอน	14,000	83	7,010	6,990
ข้อมูลแปลง	ข้อมูลทดสอบ	6,000	83	2,990	3,010

ตาราง 11 แสดงจำนวนต่าง ๆ ของชุดข้อมูลเอ็นเอสแอลเคดีดี

ประเภท		จำนวนแถว	จำนวนคุณลักษณะ	จำนวนในแต่ละประเภท	
				ปกติ	ผิดปกติ
ข้อมูลดิบ		47,735	42		
ข้อมูลสะอาด		47,735	42		
ข้อมูลแปลง	ข้อมูลฝึกสอน	25,192	118	13,449	11,743
ข้อมูลแปลง	ข้อมูลทดสอบ	22,543	118	9,710	12,833

ชุดข้อมูลมะเร็งเป็นชุดข้อมูลจากไซคิทเลิร์น และเป็นข้อมูลที่สะอาดอยู่แล้ว โดยมีจำนวนแถวทั้งหมด 569 แถว จำนวนคุณลักษณะ 30 คุณลักษณะ และแบ่งเป็น 2 ประเภท เป็นข้อมูลที่บาลานซ์แล้ว สามารถนำข้อมูลมาทำการแบ่งเป็นข้อมูลฝึกสอน และข้อมูลทดสอบ ได้โดยไม่ต้องผ่านกระบวนการทำความสะอาดข้อมูลหรือแปลงข้อมูล ซึ่งชุดข้อมูลมะเร็งนี้เป็นการทดลองข้อมูลที่ไม่เกี่ยวกับการโจมตีทางเครือข่าย ซึ่งใช้เพื่อดูความทั่วไปและดูข้อจำกัดของเฟรมเวิร์ค

ชุดข้อมูลลายมือเป็นชุดข้อมูลจากไซคิทเลิร์น เช่นเดียวกับข้อมูลชุดมะเร็ง และเป็นข้อมูลที่สะอาดอยู่แล้ว โดยมีจำนวนแถวทั้งหมด 1,797 แถว จำนวนคุณลักษณะ 64 คุณลักษณะ และแบ่งเป็น 10 ประเภท เป็นข้อมูลที่บาลานซ์แล้ว สามารถนำข้อมูลมาทำการแบ่งเป็นข้อมูลฝึกสอน และข้อมูลทดสอบ ได้โดยไม่ต้องผ่านกระบวนการทำความสะอาดข้อมูลหรือแปลงข้อมูล ซึ่งชุดข้อมูลมะเร็งนี้เป็นการทดลองข้อมูลที่ไม่เกี่ยวกับการโจมตีทางเครือข่าย ซึ่งใช้เพื่อดูความทั่วไปและดูข้อจำกัดของเฟรมเวิร์ค

3. ขั้นตอน M1 การปรับไฮเปอร์พารามิเตอร์และการหาตัวแบบที่เหมาะสม

สำหรับค่าไฮเปอร์พารามิเตอร์นั้น คือ ค่าพารามิเตอร์ต่าง ๆ ที่ผู้ใช้จะสามารถกำหนดเองก่อนที่จะทำการฝึกสอนกับชุดข้อมูลทดสอบเพื่อสร้างตัวแบบที่จะทำงานในการวิเคราะห์ข้อมูลต่าง ๆ สำหรับในขั้นตอนการปรับไฮเปอร์พารามิเตอร์ และ การหาตัวแบบแรนดอมฟอร์เรสต์ที่ดีที่สุด จุดประสงค์ของขั้นตอนนี้ คือ การค้นหาชุดของค่าไฮเปอร์พารามิเตอร์ ที่จะทำให้การทำงานของแรนดอมฟอร์เรสต์ได้ประสิทธิภาพสูงสุด โดยมีรายละเอียดการกำหนดค่าไฮเปอร์พารามิเตอร์ตามที่แสดงในตารางที่ 12 โดยรายการของค่า n_estimators และ max_depth ถูกสร้างจากการกำหนดค่าแรก ค่าสุดท้าย และจำนวนของค่าที่ต้องการ โดยที่ค่าแรกและค่าสุดท้ายอยู่ในช่วงที่ใกล้เคียงกับรายการของค่าที่ใช้สำหรับเฟรมเวิร์คการเลือกคุณลักษณะสองขั้นตอน ตัวแบบที่ผ่านการฝึกสอนตามค่าที่กำหนดจะได้รับการประเมินประสิทธิภาพกับข้อมูลทดสอบและบันทึกความแม่นยำ เพื่อเปรียบเทียบว่าค่าใดตามตาราง 12 ที่ส่งผลให้ตัวแรนดอมฟอร์เรสต์ทำงานได้ความแม่นยำสูงที่สุด

ตาราง 12 การกำหนดค่าไฮเปอร์พารามิเตอร์สำหรับขั้นตอนการปรับพารามิเตอร์ไฮเปอร์พารามิเตอร์

ชื่อพารามิเตอร์	คำอธิบาย	ค่า
n_estimators	จำนวนต้นไม้ตัดสินใจที่จะใช้ในแรนดอมฟอร์เรสต์ ประสิทธิภาพขึ้นอยู่กับจำนวนต้นไม้ แต่เวลาที่ใช้ในการสร้างตัวแบบจะนานขึ้น	[10, 36,62, 88, 114,140, 166. 192, 218, 244,270,296,322, 348, 374,400]

ชื่อพารามิเตอร์	คำอธิบาย	ค่า
max_depth	ระดับความลึกของต้นไม้ตัดสินใจ จะส่งผลต่อการจำแนกประเภทของข้อมูลได้มากยิ่งขึ้น จะหยุดการแยกโหนด ถ้าถึงระดับความลึกสูงสุดที่กำหนด	[10,26,43,60, 76, 93, 110, None]
min_samples_split	จำนวนโหนดขั้นต่ำที่จำเป็น เพื่อทำให้เกิดการแยกโหนด	[2, 4,8, 10]
min_samples_leaf	จำนวนข้อมูลขั้นต่ำใน โหนดใบ ของต้นไม้ตัดสินใจแต่ละต้น ถ้าจำนวนข้อมูลต่ำกว่าค่านี้จะหยุดการแยกโหนด	[1,2,3,4]
max features	จำนวนของคุณลักษณะที่ ต้นไม้ตัดสินใจแต่ละต้นจะสามารถใช้ในการสร้างตัวแบบ	['sqrt', 'log2']
bootstrap	ต้องการจะสุ่มต้นไม้ตัดสินใจบางส่วนของแรนดอมฟอร์เรสต์เพื่อใช้การฝึกสอนหรือไม่	[False, True]
n_iter	จำนวนรอบสูงสุดสำหรับการค้นหาค่าพารามิเตอร์ ที่เรียกการทำงานของ แรนดอมฟอร์เรสต์ ที่จะได้ค่าประสิทธิภาพสูงสุด จะหยุดการทำงานเมื่อครบรอบที่กำหนด	100
cv	ค่าที่กำหนดสำหรับการทำ Cross Validation ว่าต้องการทำกี่ Fold	5

เพื่อความเข้าใจสำหรับขั้นตอน M1 จะสรุปข้อมูลเข้า กระบวนการทำงานภายใน และผลลัพธ์ของกระบวนการดังแสดงในตาราง 12 มีชุดข้อมูลเข้าภายในจำนวนคุณลักษณะ 118 รายการ และจำนวนคุณลักษณะผลลัพธ์ 20 รายการ โดยใช้อัลกอริทึม Randomized Search CV เพื่อค้นหาค่าไฮเปอร์พารามิเตอร์

ตาราง 13 สรุปขั้นตอนการทำงาน

คุณลักษณะข้อมูลเข้า	กระบวนการ	ผลลัพธ์
ชุดข้อมูลเอ็นเอส แอลเคดีดี มีคุณลักษณะ	1.กำหนดค่าไฮเปอร์พารามิเตอร์สำหรับตัวจำแนกแรนดอมฟอร์เรสต์	1.บันทึกค่าความแม่นยำในตาราง

คุณลักษณะข้อมูลเข้า	กระบวนการ	ผลลัพธ์
ทั้งหมดที่ได้จากขั้นตอนที่ผ่านมา ได้คุณลักษณะทั้งหมดจำนวน 119 คุณลักษณะ(118 คุณลักษณะรวมกับ ทาร์เก็ต 1 คุณลักษณะ)	2.ดำเนินการปรับไฮเปอร์พารามิเตอร์โดยใช้ Randomized Search CV สุ่มเลือกค่าพารามิเตอร์ 2.1 จำนวนรอบสูงสุดสำหรับการหา ไฮเปอร์พารามิเตอร์ กำหนดตามพารามิเตอร์ n_iter ในตารางการกำหนดค่า ตั้ง n_iter ไว้ที่ 100 ฉะนั้นการทำงานจะวนซ้ำจบครบ 100 รอบ จำนวนชุดค่าผสมแบบสุ่มของไฮเปอร์พารามิเตอร์ 2.2 เลือกวิธีประเมินประสิทธิภาพตัวแบบเป็นค่าความแม่นยำ 2.3 กำหนดจำนวนสำหรับการตรวจสอบข้าม - ใช้ Randomized Search CV ในข้อมูลการฝึก ซึ่งจะฝึกและประเมินแบบจำลอง แรนดอมฟอร์เรสต์ ซ้ำๆ ด้วยการผสมผสานไฮเปอร์พารามิเตอร์ที่แตกต่างกัน - ดึงข้อมูลไฮเปอร์พารามิเตอร์ที่ดีที่สุดและตัวแบบที่เกี่ยวข้องซึ่งได้รับจาก Randomized Search CV 3.ฝึกสอนตัวจำแนกแรนดอมฟอร์เรสต์ ใหม่โดยใช้ไฮเปอร์พารามิเตอร์ที่ดีที่สุดที่ได้รับจาก Randomized Search CV และชุดข้อมูลการฝึกสอนทั้งหมด 4.ประเมินประสิทธิภาพของตัวแบบที่ผ่านการฝึกสอนจากข้อมูลการทดสอบเพื่อประเมินความแม่นยำ	ผลลัพธ์การทดลอง 2.ได้ไฮเปอร์พารามิเตอร์ที่ดีที่สุด 3.ได้คุณลักษณะที่มีความสัมพันธ์กับเป้าหมายมากที่สุดตามจำนวนลักษณะที่กำหนด

เมื่อดำเนินการทำงานค้นหาค่าไฮเปอร์พารามิเตอร์ที่ดีที่สุดตามค่าที่กำหนดในตาราง 13 ด้วยการใช้กระบวนการ Random Search CV ที่มีขั้นตอนการทำงานในตาราง จะได้ผลคือค่าพารามิเตอร์ตามที่แสดงในตาราง ซึ่งในส่วนของผลลัพธ์จะประกอบด้วย ค่าความแม่นยำสูงสุดที่

ทำได้คือ 0.7751 และชื่อค่าคุณลักษณะที่ดีที่สุดจำนวน 20 รายการ และค่าไฮเปอร์พารามิเตอร์ที่ทำให้ได้ค่าดังกล่าว

ตาราง 14 แสดงตัวอย่างการหนดค่าไฮเปอร์พารามิเตอร์ที่ให้ผลการทำงานดีที่สุดสำหรับ M1

คุณลักษณะข้อมูล	กระบวนการ	ผลลัพธ์
คุณลักษณะทาร์เก็ตรวม เป็น 119 คุณลักษณะ กำหนดจำนวนคุณลักษณะที่ต้องการ 20 คุณลักษณะ	1.กำหนดค่าไฮเปอร์พารามิเตอร์สำหรับตัวจำแนกแรนดอมฟอร์เรสต์ 2.ดำเนินการปรับไฮเปอร์พารามิเตอร์ 3.ฝึกสอนตัวจำแนกแรนดอมฟอร์เรสต์ ใหม่โดยใช้ไฮเปอร์พารามิเตอร์ที่ดีที่สุดที่ได้รับจาก Randomized Search CV และชุดข้อมูลการฝึกสอนทั้งหมด 4.ประเมินประสิทธิภาพของตัวแบบที่ผ่านการฝึกสอนจากข้อมูลการทดสอบเพื่อประเมินความแม่นยำ	1.Accuracy (Train Accuracy = 1.00 ,Test Accuracy = 0.7751 2.คุณลักษณะ (FI) 'dst_bytes', 'src_bytes', 'diff_srv_rate', 'same_srv_rate', 'dst_host_same_srv_rate', 'dst_host_srv_count', 'logged_in', 'flag_SF', 'dst_host_diff_srv_rate', 'serror_rate', 'srv_serror_rate', 'dst_host_same_src_port_rate', 'flag_S0', 'service_http', 'count', 'srv_count', 'protocol_type_icmp', 'dst_host_serror_rate', 'dst_host_count', 'dst_host_srv_rerror_rate' 3.Hyperพารามิเตอร์ 'n_estimators': 374, 'min_samples_split': 4, 'min_samples_leaf': 1, 'max_features': 'sqrt’, 'max_depth': 43, 'bootstrap': False

4. ขั้นตอน การเลือกคุณลักษณะขั้นตอนที่ 1 (1st Stage Feature Select)

จุดประสงค์ของขั้นตอนแรกของการเลือกคุณลักษณะคือการลดขนาดของชุดข้อมูล เนื่องจากการสร้างคุณสมบัติพหุนามจากคุณสมบัติเดิมจะเพิ่มจำนวนคุณลักษณะเพิ่มขึ้นอย่างมาก ส่งผลกระทบทรัพยากรในการประมวลผลที่มากขึ้นตามไปด้วย ดังนั้นในขั้นตอนนี้มีจุดประสงค์คือ เลือกคุณลักษณะเกี่ยวข้องมากที่สุดซึ่งมีความสัมพันธ์สูงสุดกับตัวแปรเอาต์พุต วิธีการเลือกคุณลักษณะ ประกอบด้วยวิธีไคสแควร์, ค่าการวิเคราะห์ความแปรปรวน, สารสนเทศร่วม และความสำคัญของคุณลักษณะ วิธีการเหล่านี้จะทำการคำนวณค่าคะแนนให้กับคุณลักษณะแต่ละรายการ จากนั้นทำการเรียงค่าคุณลักษณะตามค่าที่ได้คะแนนจากแต่ละวิธีและเลือกคุณลักษณะจำนวน k อันดับแรก ในการศึกษานี้ เลือก k จากเซต {20, 30,40}

แต่ละคุณลักษณะจะได้รับการประเมินแยกกันโดยใช้การวัดทางสถิติหรือฟังก์ชันการให้คะแนนเพื่อกำหนดความเกี่ยวข้องกับตัวแปรเป้าหมาย

- ไคสแควร์ คือกำหนดความเป็นอิสระระหว่างคุณลักษณะตามหมวดหมู่และตัวแปรเป้าหมาย

- การวิเคราะห์ความแปรปรวน คือประเมินนัยสำคัญทางสถิติของความแตกต่างของค่าเฉลี่ยระหว่างกลุ่มที่กำหนดโดยคุณลักษณะที่เป็นหมวดหมู่

- สารสนเทศร่วม คือคำนวณการพึ่งพาซึ่งกันและกันระหว่างคุณสมบัติและตัวแปรเป้าหมาย

- ความสำคัญของคุณลักษณะ คือคุณลักษณะที่ผ่านการหาตัวแบบที่เหมาะสม จะได้คุณลักษณะที่มีความสำคัญกับตัวแบบ

โดยวิธีการดังกล่าวนี้ ถูกจัดในกลุ่มที่เรียกว่า การกรองคุณลักษณะ

ตาราง 15 แสดงขั้นตอนทำงาน

ขั้นตอนที่	คำอธิบาย
1	ใช้เทคนิคการเลือกคุณลักษณะไคสแควร์เพื่อคำนวณสถิติไคสแควร์สำหรับแต่ละคุณลักษณะและตัวแปรเป้าหมาย
2	เลือกคุณลักษณะ K อันดับต้นที่มีคะแนนไคสแควร์สูงสุด โดยที่ K คือตัวเลขที่กำหนดไว้ล่วงหน้าหรือเปอร์เซ็นต์ของจำนวนคุณลักษณะทั้งหมด
3	ใช้เทคนิคการเลือกคุณลักษณะการวิเคราะห์ความแปรปรวน เพื่อวัดค่า F

ขั้นตอนที่	คำอธิบาย
	ระหว่างแต่ละคุณลักษณะและตัวแปรเป้าหมาย
4	เลือกคุณสมบัติ K อันดับต้นที่มีค่า F สูงสุด โดยที่ K คือตัวเลขที่กำหนดไว้ล่วงหน้าหรือเปอร์เซ็นต์ของจำนวนคุณสมบัติทั้งหมด
5	ใช้เทคนิคการเลือกคุณลักษณะของาสารสนเทศร่วมเพื่อประเมินข้อมูลร่วมกันระหว่างแต่ละคุณลักษณะและตัวแปรเป้าหมาย
6	เลือกคุณลักษณะ K อันดับต้นที่มีคะแนนข้อมูลร่วมกันสูงสุด โดยที่ K คือตัวเลขที่กำหนดไว้ล่วงหน้าหรือเปอร์เซ็นต์ของจำนวนคุณลักษณะทั้งหมด
7	ใช้เทคนิคความสำคัญของคุณลักษณะ เช่น การใช้ตัวแยกประเภทแรนดอมฟอเรสต์ เพื่อกำหนดความสำคัญของคุณลักษณะแต่ละอย่างตามการมีส่วนร่วมกับประสิทธิภาพของแบบจำลอง
8	เลือกคุณลักษณะ K อันดับต้นที่มีคะแนนความสำคัญของคุณลักษณะสูงสุด โดยที่ K คือตัวเลขที่กำหนดไว้ล่วงหน้าหรือเปอร์เซ็นต์ของจำนวนคุณลักษณะทั้งหมด
9	รวมคุณสมบัติที่เลือกจากเทคนิคทั้งหมดเป็นชุดคุณสมบัติสุดท้ายและฝึกสอนแบบจำลอง
10	ฝึกตัวแบบโดยใช้คุณสมบัติที่เลือกในชุดการฝึก
11	ประเมินประสิทธิภาพความแม่ยำของแบบจำลองในชุดทดลอบ

5. ขั้นตอน M2 การปรับไฮเปอร์พารามิเตอร์และการปรับตัวแบบ

ขั้นตอนที่ 5 มีการทำงานเหมือนกับขั้นตอนที่ 3 รวมถึงรายละเอียดการตั้งค่าไฮเปอร์พารามิเตอร์จะเหมือนกัน ดังนั้นในส่วนนี้จะไม่ทำการอธิบายส่วนที่เหมือนกัน แต่จะอธิบายถึงผลลัพธ์จากการทำงานดังแสดงในตาราง aa4 ขั้นตอน การเลือกคุณลักษณะขั้นตอนที่ 1 ในขั้นตอนนี้จะทำงานเหมือนขั้นตอน M1 ดังนี้

ตาราง 16 แสดงตัวอย่างการหนดค่าไฮเปอร์พารามิเตอร์ที่ให้ผลการทำงานดีที่สุดสำหรับ M2

คุณลักษณะข้อมูล	กระบวนการ	ผลลัพธ์
กำหนดจำนวนคุณลักษณะที่ต้องการ 20 คุณลักษณะ	1.กำหนดค่าไฮเปอร์พารามิเตอร์สำหรับตัวจำแนกแรนดอมฟอร์เรสต์ 2.ดำเนินการปรับไฮเปอร์พารามิเตอร์ 3.ฝึกสอนตัวจำแนกแรนดอมฟอร์เรสต์ ใหม่โดยใช้ไฮเปอร์พารามิเตอร์ที่ดีที่สุดที่ได้รับจาก Randomized Search CV และชุดข้อมูลการฝึกสอนทั้งหมด 4.ประเมินประสิทธิภาพของตัวแบบที่ผ่านการฝึกสอนจากข้อมูลการทดสอบเพื่อประเมินความแม่นยำ	1.Accuracy (Train Accuracy = 1.00 ,Test Accuracy = 0.769 2.ไฮเปอร์พารามิเตอร์ 'n_estimators': 192, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_features': 'log2', 'max_depth': None, 'bootstrap': False

6.ขั้นตอนการสร้างคุณลักษณะพหุนาม

ในขั้นตอนนี้ สร้างคุณสมบัติพหุนามจากคุณลักษณะที่เลือกโดยใช้วิธีสร้างคุณสมบัติพหุนาม จุดประสงค์ของขั้นตอนนี้คือการสร้างคุณลักษณะเพิ่มเติมที่มีลำดับสูงกว่าระหว่างคุณลักษณะที่เลือก ใช้การสร้างคุณสมบัติพหุนามกับคุณสมบัติที่เลือกในขั้นตอนก่อนหน้าของการเลือกคุณสมบัติ ซึ่งใช้วิธีอิงตามตัวกรอง วิธีการสร้างคุณสมบัติพหุนามเกี่ยวข้องกับการสร้างคุณสมบัติใหม่โดยการคำนวณที่เป็นไปได้ทั้งหมดของคุณสมบัติที่เลือกจนถึงระดับหนึ่ง ตัวอย่างเช่น ถ้ามีสองคุณลักษณะ X1 และ X2 การสร้างคุณลักษณะพหุนามดีกรี 2 จะสร้างคุณลักษณะใหม่ เช่น X1^2, X2^2 และ X1*X2

ระดับของการสร้างคุณสมบัติพหุนามจะกำหนดจำนวนของคุณสมบัติใหม่ที่สร้างขึ้น ดีกรีพหุนามที่สูงขึ้นส่งผลให้มีคุณลักษณะมากขึ้น แต่ยังเพิ่มความเสี่ยงของการโอเวอร์ฟิตติ้งและความซับซ้อนในการคำนวณด้วย ในการศึกษานี้ ใช้การสร้างคุณสมบัติพหุนามดีกรี 2 และ 3 ผลลัพธ์ของ

ขั้นตอนนี้คือชุดคุณลักษณะใหม่ที่รวมคุณลักษณะเดิมที่เลือกไว้ ตลอดจนคุณลักษณะพหุนามที่สร้างขึ้นใหม่ คุณลักษณะเหล่านี้จะใช้ในขั้นตอนต่อไปของการปรับไฮเปอร์พารามิเตอร์และการปรับตัวแบบ

ตาราง 17 แสดงตัวอย่างการหนดค่าไฮเปอร์พารามิเตอร์ที่ให้ผลการทำงานดีที่สุดสำหรับ M2

ดีกรี	จำนวนคุณลักษณะ (ก่อน)	จำนวนคุณลักษณะ (หลัง)
1	20	20
2	20	210
3	20	1,350

7. ขั้นตอน M3 การปรับไฮเปอร์พารามิเตอร์และการปรับตัวแบบ

ในขั้นตอนนี้ ทำการปรับไฮเปอร์พารามิเตอร์ และการปรับตัวแบบโดยใช้คุณสมบัติที่สร้างขึ้นจากขั้นตอนก่อนหน้าของการสร้างคุณสมบัติคุณลักษณะ จุดประสงค์ของขั้นตอนนี้ คือ การค้นหาชุดของไฮเปอร์พารามิเตอร์ที่ดีที่สุดสำหรับตัวแยกประเภทของแรนดอมฟอเรสต์ และฝึกตัวแบบที่ดีที่สุดโดยใช้ไฮเปอร์พารามิเตอร์เหล่านี้ ตัวแบบที่ผ่านการฝึกสอนจะได้รับการประเมินจากข้อมูลการทดสอบเพื่อให้ได้ความแม่นยำของตัวแบบ

ตาราง 18 แสดงตัวอย่างคุณลักษณะตามขั้นตอนการปรับพารามิเตอร์ไฮเปอร์พารามิเตอร์และการหาตัวแบบที่เหมาะสม

คุณลักษณะข้อมูล	กระบวนการ	ผลลัพธ์

จำนวนคุณลักษณะที่ผ่านกระบวนการขั้นตอนการสร้างคุณลักษณะพหุนามกำหนดจำนวนคุณลักษณะที่ต้องการ 20 คุณลักษณะ	1.กำหนดค่าไฮเปอร์พารามิเตอร์สำหรับตัวจำแนกแรนดอมฟอร์เรสต์ 2.ดำเนินการปรับไฮเปอร์พารามิเตอร์ 3.ฝึกสอนตัวจำแนกแรนดอมฟอร์เรสต์ ใหม่โดยใช้ไฮเปอร์พารามิเตอร์ที่ดีที่สุดที่ได้รับจาก Randomized Search CV และชุดข้อมูลการฝึกสอนทั้งหมด 4.ประเมินประสิทธิภาพของตัวแบบที่ผ่านการฝึกสอนจากข้อมูลการทดสอบเพื่อประเมินความแม่นยำ	1.ความแม่นยำ (Train Accuracy = 1.00 ,Test Accuracy = 0.781 2.ไฮเปอร์พารามิเตอร์ {'n_estimators': 192, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_features': 'log2', 'max_depth': None, 'bootstrap': False}

8. ขั้นตอน การเลือกคุณลักษณะขั้นตอนที่ 2

การเลือกคุณลักษณะขั้นที่ 2 มีจุดประสงค์เพื่อปรับแต่งชุดคุณลักษณะเพิ่มเติมโดยการลบความสัมพันธ์ของคุณลักษณะที่ไม่จำเป็นซึ่งสร้างในขั้นตอนก่อนหน้า ด้วยเหตุนี้ ตารางจึงแสดงคุณลักษณะที่สอดคล้องกันหลังจากกระบวนการเลือกคุณลักษณะ

ตาราง 19 แสดงขั้นตอนการทำงาน

ขั้นตอนที่	คำอธิบาย
1	ใช้เทคนิคการเลือกคุณลักษณะไคสแควร์เพื่อคำนวณสถิติไคสแควร์สำหรับแต่ละคุณลักษณะและตัวแปรเป้าหมาย
2	เลือกคุณลักษณะ K อันดับต้นที่มีคะแนนไคสแควร์สูงสุด โดยที่ K คือตัวเลขที่กำหนดไว้ล่วงหน้าหรือเปอร์เซ็นต์ของจำนวนคุณลักษณะทั้งหมด
3	ใช้เทคนิคการเลือกคุณลักษณะการวิเคราะห์ความแปรปรวน เพื่อวัดค่า F ระหว่างแต่ละคุณลักษณะและตัวแปรเป้าหมาย
4	เลือกคุณสมบัติ K อันดับต้นที่มีค่า F สูงสุด โดยที่ K คือตัวเลขที่กำหนดไว้ล่วงหน้าหรือเปอร์เซ็นต์ของจำนวนคุณสมบัติทั้งหมด
5	ใช้เทคนิคการเลือกคุณลักษณะของงาสารสนเทศร่วมเพื่อประเมินข้อมูลร่วมกันระหว่างแต่ละคุณลักษณะและตัวแปรเป้าหมาย

ขั้นตอนที่	คำอธิบาย
6	เลือกคุณลักษณะ K อันดับต้นที่มีคะแนนข้อมูลร่วมกันสูงสุด โดยที่ K คือตัวเลขที่กำหนดไว้ล่วงหน้าหรือเปอร์เซ็นต์ของจำนวนคุณลักษณะทั้งหมด
7	ใช้เทคนิคความสำคัญของคุณลักษณะ เช่น การใช้ตัวแยกประเภทแรนดอมฟอเรสต์ เพื่อกำหนดความสำคัญของคุณลักษณะแต่ละอย่างตามการมีส่วนร่วมกับประสิทธิภาพของตัวแบบ
8	เลือกคุณลักษณะ K อันดับต้นที่มีคะแนนความสำคัญของคุณลักษณะสูงสุด โดยที่ K คือตัวเลขที่กำหนดไว้ล่วงหน้าหรือเปอร์เซ็นต์ของจำนวนคุณลักษณะทั้งหมด
9	รวมคุณสมบัติที่เลือกจากเทคนิคทั้งหมดเป็นชุดคุณสมบัติสุดท้ายและฝึกสอนตัวแบบ
10	ฝึกตัวแบบโดยใช้คุณสมบัติที่เลือกในชุดข้อมูลการฝึกสอน
11	ประเมินประสิทธิภาพความแม่ยำของแบบจำลองในชุดข้อมูลทดสอบ

9. ขั้นตอน M4 การปรับไฮเปอร์พารามิเตอร์และการปรับตัวแบบ

ในขั้นตอนนี้ ทำการปรับไฮเปอร์พารามิเตอร์ และการปรับตัวแบบโดยใช้คุณลักษณะที่เลือกซึ่งได้รับจากการเลือกคุณลักษณะในขั้นตอนพหุนาม จุดประสงค์ของขั้นตอนนี้คือการค้นหาชุดของไฮเปอร์พารามิเตอร์ที่ดีที่สุดสำหรับตัวแยกประเภทแรนดอมฟอเรสต์โดยใช้คุณลักษณะที่เลือกและฝึกตัวแบบที่ดีที่สุดโดยใช้ไฮเปอร์พารามิเตอร์เหล่านี้ ตัวแบบที่ผ่านการฝึกสอนจะได้รับการประเมินจากข้อมูลการทดสอบเพื่อให้ได้ความแม่นยำของตัวแบบ

ตาราง 20 แสดงตัวอย่างคุณลักษณะตามขั้นตอนการปรับพารามิเตอร์ไฮเปอร์พารามิเตอร์และการหาตัวแบบที่เหมาะสม

คุณลักษณะข้อมูล	กระบวนการ	ผลลัพธ์

จำนวนคุณลักษณะที่ผ่านกระบวนการขั้นตอนการสร้างคุณลักษณะพหุนาม กำหนดจำนวนคุณลักษณะที่ต้องการ 20 คุณลักษณะ	1.กำหนดค่าไฮเปอร์พารามิเตอร์สำหรับตัวจำแนกแรนดอมฟอร์เรสต์ 2.ดำเนินการปรับไฮเปอร์พารามิเตอร์ 3.ฝึกสอนตัวจำแนกแรนดอมฟอร์เรสต์ ใหม่โดยใช้ไฮเปอร์พารามิเตอร์ที่ดีที่สุดที่ได้รับจาก Randomized Search CV และชุดข้อมูลการฝึกสอนทั้งหมด 4.ประเมินประสิทธิภาพของตัวแบบที่ผ่านการฝึกสอนจากข้อมูลการทดสอบเพื่อประเมินความแม่นยำ	1.ความแม่นยำ (Train Accuracy = 1.00 ,Test Accuracy = 0.778 2.คุณลักษณะ (FI) 'src_bytes', 'dst_bytes', 'same_srv_rate', 'flag_SF', 'diff_srv_rate', 'dst_host_same_srv_rate', 'count', 'dst_host_srv_count', 'dst_host_diff_srv_rate', 'logged_in', 'srv_serror_rate', 'dst_host_same_src_port_rate', 'dst_host_srv_serror_rate', 'dst_host_serror_rate', 'protocol_type_icmp', 'service_http', 'serror_rate', 'dst_host_srv_diff_host_rate', 'flag_S0', 'dst_host_count' 3.Hyperพารามิเตอร์ {'n_estimators': 374, 'min_samples_split': 4, 'min_samples_leaf': 1, 'max_features': 'sqrt', 'max_depth': 43, 'bootstrap': False}

10. ตารางผลลัพธ์

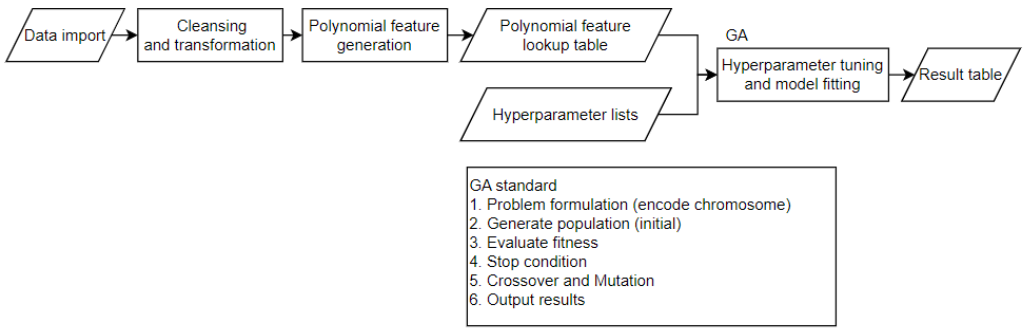
ในขั้นตอนนี้ สร้างตารางเพื่อสรุปผลลัพธ์ที่ได้รับในแต่ละขั้นตอนของเฟรมเวิร์คสำหรับชุดข้อมูลแต่ละชุด และสำหรับการดำเนินการของเฟรมเวิร์คแต่ละรายการ จุดประสงค์ของขั้นตอนนี้คือเพื่อเปรียบเทียบความแม่นยำที่ได้รับในแต่ละขั้นตอน และเพื่อเลือกชุดไฮเปอร์พารามิเตอร์และคุณสมบัติที่ดีที่สุดสำหรับชุดข้อมูลแต่ละชุด

3.2.2 ขั้นตอนวิธีการคัดเลือกคุณลักษณะข้อมูลที่ใช้เทคนิคพหุนามร่วมกับขั้นตอนวิธีเชิงพันธุกรรม

ตามวัตถุประสงค์วิทยานิพนธ์ข้อที่ 2 เพื่อพัฒนาขั้นตอนวิธีสำหรับการเลือกคุณสมบัติและการสร้างคุณสมบัติแบบพหุนามด้วยขั้นตอนวิธีเชิงพันธุกรรม และ เพื่อประยุกต์ใช้เฟรมเวิร์คหรือขั้นตอนวิธีเพื่อวัตถุประสงค์ในการลดพื้นที่การจัดเก็บเหตุการณ์ที่เกิดขึ้นในระบบเครือข่าย วัตถุประสงค์วิทยานิพนธ์ข้อที่ 2 จุดมุ่งหมายคือการพัฒนาอัลกอริทึมที่ใช้อัลกอริธึมเชิงพันธุกรรม

สำหรับการเลือกคุณลักษณะและการสร้างคุณลักษณะพหุนาม อัลกอริทึมทางพันธุกรรมเป็นเทคนิคการเพิ่มประสิทธิภาพการคำนวณ สามารถใช้เพื่อค้นหาโซลูชันที่เหมาะสมที่สุดภายในพื้นที่โซลูชันขนาดใหญ่ การเลือกคุณลักษณะคือกระบวนการระบุคุณลักษณะที่เกี่ยวข้องและให้ข้อมูลมากที่สุดจากชุดข้อมูลที่กำหนด การใช้อัลกอริธึมเชิงพันธุกรรม มีวัตถุประสงค์เพื่อออกแบบอัลกอริทึมที่สามารถเลือกชุดย่อยของคุณลักษณะที่สนับสนุนพลังการทำนายของแบบจำลองได้มากที่สุดโดยอัตโนมัติ อัลกอริทึมเชิงพันธุกรรมจะประเมินชุดค่าผสมของคุณลักษณะต่างๆ พัฒนาผ่านการวนซ้ำและเลือกชุดย่อยของคุณลักษณะที่ดีที่สุดตามประสิทธิภาพในเมตริกการประเมินเฉพาะ นอกจากนี้การสร้างคุณสมบัติพหุนามยังเกี่ยวข้องกับการสร้างคุณสมบัติใหม่โดยการรวมคุณสมบัติที่มีอยู่โดยใช้ฟังก์ชันพหุนาม สิ่งนี้สามารถช่วยจับความสัมพันธ์ที่ไม่ใช่เชิงเส้นระหว่างตัวแปรและอาจปรับปรุงประสิทธิภาพของตัวแบบการเรียนรู้ด้วยตัวเองของคอมพิวเตอร์ อัลกอริธึมเชิงพันธุกรรมจะสำรวจชุดค่าผสมพหุนามต่าง ๆ เลือกคุณสมบัติพหุนามที่มีค่าที่สุด และละทิ้งคุณสมบัติซ้ำซ้อนหรือไม่เกี่ยวข้อง วัตถุประสงค์วิทยานิพนธ์ข้อที่ 3 การใช้เฟรมเวิร์กหรืออัลกอริทึมเพื่อลดพื้นที่จัดเก็บสำหรับเหตุการณ์ในระบบเครือข่าย ระบบเครือข่ายสร้างข้อมูลเหตุการณ์จำนวนมาก และการจัดเก็บเหตุการณ์ทั้งหมดอาจใช้ทรัพยากรมาก ดังนั้น การค้นหาวิธีที่มีประสิทธิภาพเพื่อลดพื้นที่จัดเก็บในขณะที่รักษาข้อมูลที่สำคัญจึงเป็นสิ่งสำคัญ การรวมเหตุการณ์ที่คล้ายคลึงกันเป็นการนำเสนอโดยสรุปหรือรูปแบบระดับที่สูงขึ้น ซึ่งอาจเกี่ยวข้องกับการจัดกลุ่มเหตุการณ์ตามคุณลักษณะเฉพาะ ช่วงเวลา หรือคุณสมบัติทั่วไปเพื่อลดพื้นที่จัดเก็บที่จำเป็น การแยกคุณลักษณะที่เกี่ยวข้องหรือข้อมูลสำคัญจากเหตุการณ์และจัดเก็บเฉพาะคุณลักษณะเหล่านั้นแทนการบันทึกเหตุการณ์ทั้งหมด ซึ่งอาจเกี่ยวข้องกับเทคนิคต่างๆ เช่น การวิเคราะห์องค์ประกอบหลัก หรือการแฮชคุณลักษณะ

ทางเลือกของขั้นตอนวิธีหรืออัลกอริทึมสำหรับการลดพื้นที่จัดเก็บจะขึ้นอยู่กับข้อกำหนดเฉพาะของระบบเครือข่าย ประเภทของเหตุการณ์ และการแลกเปลี่ยนที่ต้องการระหว่างประสิทธิภาพการจัดเก็บและการเก็บรักษาข้อมูล ประกอบด้วยขั้นตอนดังต่อไปนี้



ภาพ 5 ขั้นตอนวิธีการคัดเลือกคุณลักษณะข้อมูลที่ใช้เทคนิคพหุนามร่วมกับขั้นตอนวิธีเชิงพันธุกรรม

ขั้นตอนวิธีการคัดเลือกคุณลักษณะข้อมูลที่ใช้เทคนิคพหุนามร่วมกับขั้นตอนวิธีเชิงพันธุกรรม

1. ขั้นตอน การนำเข้าชุดข้อมูล คือ กระบวนการนำเข้าชุดข้อมูลที่มีข้อมูลการบุกรุกเครือข่าย

2. ขั้นตอน การทำความสะอาดชุดข้อมูลและแปลงข้อมูล คือ การทำความสะอาดข้อมูลและแปลงข้อมูลคือทำความสะอาดและประมวลผลข้อมูลล่วงหน้า จัดการค่าที่ขาดหายไป และแปลงตัวแปรหมวดหมู่เป็นตัวเลข

3. ขั้นตอน การสร้างคุณลักษณะพหุนาม คือ สร้างคุณสมบัติพหุนามโดยใช้คุณสมบัติที่เลือกจากชุดข้อมูลต้นฉบับ

4. ขั้นตอนบันทึกตารางคุณลักษณะพหุนาม คือ สร้างตารางค้นหาเพื่อจัดเก็บคุณสมบัติพหุนามที่สร้างขึ้นและคุณสมบัติเดิมที่สอดคล้องกัน ตารางนี้จะใช้เพื่อจับคู่คุณสมบัติพหุนามที่เลือกกลับไปยังคุณสมบัติเดิม

5. ขั้นตอนไฮเปอร์พารามิเตอร์ลิสต์ คือกำหนดรายการของไฮเปอร์พารามิเตอร์ที่จะใช้สำหรับอัลกอริธึมเชิงพันธุกรรมและแบบจำลองแรนดอมฟอเรสต์ ตัวอย่าง ได้แก่ n_estimators, max_depth, min_samples_split, max_features, min_samples_leaf และ bootstrap โดยรายการของค่า n_estimators และ max_depth ถูกสร้างจากการกำหนดค่าเริ่มต้น ค่าสุดท้าย และจำนวนค่าที่ต้องการ

ตาราง 21 แสดงเซตของไฮเปอร์พารามิเตอร์

พารามิเตอร์	คำอธิบาย	ค่า
n_estimators	จำนวนต้นไม้การตัดสินใจในต้นไม้ตัดสินใจ	[10, 36,62, 88, 114,140, 166. 192, 218, 244,270,296,322, 348, 374,400]
max_depth	ความลึกสูงสุด สำหรับแผนผังการตัดสินใจแต่ละรายการในชุด	[10,26,43,60, 76, 93, 110, None]
min_samples_split	จำนวนตัวอย่างขั้นต่ำที่จำเป็นในการดำเนินการแยกที่โหนดในแผนผังการ	[2, 4,8, 10]

พารามิเตอร์	คำอธิบาย	ค่า
	ตัดสินใจ	
min_samples_leaf	จำนวนของคุณสมบัติที่ต้องพิจารณาเมื่อค้นหาการแยกที่ดีที่สุดในแต่ละโหนดของต้นไม้	[1,2,3,4]
min_samples_leaf	จำนวนตัวอย่างขั้นต่ำที่จำเป็นสำหรับโหนดปลายสุดในแผนผังการตัดสินใจ	['sqrt', 'log2']
bootstrap	ต้องการจะสุ่มต้นไม้ตัดสินใจบางส่วนของแรนดอมฟอร์เรสต์เพื่อใช้การฝึกสอนหรือไม่	[False, True]
degree	จำนวนดีกรีสำหรับการสร้างคุณลักษณะพหุนาม	[1, 2, 3, 4]

6. การปรับไฮเปอร์พารามิเตอร์และการหาตัวแบบที่เหมาะสม ขั้นตอน ใช้อัลกอริธึมเชิงพันธุกรรมเพื่อค้นหาส่วนผสมที่ดีที่สุดของคุณสมบัติพหุนามและไฮเปอร์พารามิเตอร์ แบบจำลองแรนดอมฟอเรสต์ด้วยคุณสมบัติพหุนามและไฮเปอร์พารามิเตอร์ที่เลือก ประเมินประสิทธิภาพของแบบจำลองโดยใช้ชุดฝึกสอนและชุดทดสอบ ในกรอบแนวคิดนี้ใช้วิธีเชิงพันธุกรรม เพื่อทำความเข้าใจในกระบวนการของขั้นตอนวิธีเชิงพันธุกรรมทั้งหมด ทำให้ทราบว่ามีฟังก์ชันต่าง ๆ ดังนี้

1. ฟังก์ชันการสร้างต้นแบบพันธุกรรม ทำหน้าที่สุ่มสร้างแต่ละโครโมโซม โดยทำการสร้างทีละ 1 ยีน ซึ่งแต่ละยีนจะต้องไม่มียีนใดซ้ำกัน แต่ในโครโมโซมสามารถซ้ำกันได้เนื่องจากเป็นการเลียนแบบธรรมชาติ การสร้างประชากร โดย เอาคุณลักษณะต่าง ๆ ของคำตอบที่เป็นไปได้มาทำให้อยู่ในรูปแบบโครโมโซม จะสร้างโครโมโซม เป็น ความยาวตามจำนวนของคุณลักษณะของชุดข้อมูลและค่าพารามิเตอร์ของแรนดอม ยีน ยกตัวอย่างเช่น ในกรณีที่จำนวนคุณลักษณะมี ค่าเท่ากับ 20 คุณลักษณะ ดังนั้นโครโมโซมจึงประกอบด้วย

ยีน 1-20 คือ จำนวนคุณลักษณะที่เข้ามา

ยีน 21-22 คือ พารามิเตอร์ ดีกรี ของพหุนาม

ยีน 23-27 คือ พารามิเตอร์ n_estimators ของ แรนดอมฟอร์เรสต์

ยีน 28-30 คือ พารามิเตอร์ max_depth ของ แรนดอมฟอร์เรสต์

ยีน 31-32 คือ พารามิเตอร์ min_samples_split ของ แรนดอมฟอร์เรสต์

ยีน 33 คือ พารามิเตอร์ max_features ของ แรนดอมฟอร์เรสต์

ยีน 34-6 คือ พารามิเตอร์ max_samples ของ แรนดอมฟอร์เรสต์

ยีน 37 คือ พารามิเตอร์ bootstrap ของ แรนดอมฟอร์เรสต์

เช่น Chromosome = Gene1+Gene2+Gene3+.....+Gene37

2. ขั้นตอน Fitness Measure คำนวณหาค่าลูกที่มีประสิทธิภาพในการผสมพันธ์และการกลายพันธ์เพื่อให้ได้ค่าที่เหมาะสมที่สุดเอาไว้ มีขั้นตอนดังนี้

- ขั้นตอนถอดรหัสโครโมโซม
- ขั้นตอนการเลือกคุณลักษณะโดยทำการเลือก เฉพาะ คุณลักษณะที่มีค่าเป็น 1 เช่น 0 1 0 1 1 0 1 0 1 1 1 0 1 0 1 0 1 1 0 1
- ขั้นตอนนำคุณลักษณะที่ได้เลือก มาแปลงเป็น พหุนาม
- ขั้นตอนหาไฮเปอร์พารามิเตอร์ โดยขั้นตอนวิธีเชิงพันธุกรรม
- นำค่าคุณลักษณะที่ได้ไปค้นหา ค่าจากใน Polynomial feature lookup table
- นำค่าคุณลักษณะ ที่ได้และ พารามิเตอร์ ทำการพยากรณ์กับแรนดอมฟอร์เรสต์อัลกอริธึม แล้ว เก็บผลลัพธ์จากการพยากรณ์
- ตรวจสอบค่าจำนวนรอบ ว่าเท่ากับ รุ่น ที่ตั้ง หากว่ายังให้ส่งพ่อแม่ชุดใหม่กลับไปเข้ากรรมวิธีเชิงพันธุกรรมอีกครั้ง ถ้าจำนวนรอบเท่า รุ่น ที่ตั้งไว้ ให้ข้ามขั้นตอนนี้ โดยนำค่าความแข็งแรงของโครโมโซมมากำหนดความน่าจะเป็นให้แต่ละโคโมโซมในขั้นตอนนี้ จะทำให้ได้โครโมโซมที่แข็งแรงมากกว่าโคโมโซมที่อ่อนแอ และสุ่มเลือกประชากรจำนวนเท่ากับขนาดของประชากรเก็บค่าไว้เป็นพ่อแม่ชุดใหม่
- ฟังก์ชัน Genetic Operation ทำหน้าที่เลียนแบบกระบวนการสร้างประชากรใหม่ ในแต่ละรุ่น โดยแบ่งออกเป็น 2 กระบวนการคือ การสลับสายพันธุ์ และ Mutation
- ฟังก์ชันการคำนวณค่าผลลัพธ์ของโครโมโซมเพื่อนำไปวัดค่าความแข็งแรงของแต่ละโครโมโซม
- ฟังก์ชัน Roulette Wheel ทำหน้าที่คัดเลือกโครโมโซม โดยดูจากความแข็งแรงของโครโมโซมเป็นตัวกำหนดความน่าจะเป็นที่จะได้รับการคัดเลือกของแต่ละโครโมโซม
- ฟังก์ชัน Chromosome Selection การสรรหาโครโมโซม ทำหน้าที่สุ่มเลือกโครโมโซมเข้ากระบวนการ Genetic Operation

7. ตารางผลลัพธ์ ขั้นตอน รวบรวมผลลัพธ์ของอัลกอริธึมเชิงพันธุกรรมและแบบจำลองแรนดอมฟอเรสต์ลงในตาราง ตารางควรประกอบด้วยการสร้าง ระดับ จำนวนคุณสมบัติ จำนวนของคุณสมบัติพหุนาม ไฮเปอร์พารามิเตอร์ที่เลือก และความแม่นยำในการฝึกและทดสอบ จัดเรียงตารางตามความแม่นยำในการทดสอบจากมากไปน้อยเพื่อระบุรุ่นที่มีประสิทธิภาพดีที่สุด

ลำดับขั้นตอนในการทดลอง

1.การนำเข้าชุดข้อมูล

ใช้ชุดข้อมูลที่ได้ทำตามวัตุประสงค์ข้อที่ 1 มาใช้งานในขั้นตอนนี้ได้ เพื่อลดเวลาในการทำงาน

2.การทำความสะอาดชุดข้อมูลและการแปลงข้อมูล

ใช้ชุดข้อมูลที่ได้ทำตามวัตุประสงค์ข้อที่ 1 มาใช้งานในขั้นตอนนี้ได้ เพื่อลดเวลาในการทำงาน

3. การสร้างคุณลักษณะพหุนาม เพื่อสร้างความสัมพันธ์ระหว่างคุณลักษณะแบบไม่เป็นเชิงเส้น

4. ตารางบันทึกคุณลักษณะพหุนาม คือ สร้างตารางค้นหาเพื่อจัดเก็บคุณสมบัติพหุนามที่สร้างขึ้นและคุณสมบัติเดิมที่สอดคล้องกัน

4.1 หลังจากสร้างคุณสมบัติพหุนามในขั้นตอนที่ 3 แล้ว ความสัมพันธ์ระหว่างคุณสมบัติที่สร้างขึ้นใหม่และตัวแปรเดิม คุณสมบัติพหุนามแต่ละรายการคือการรวมกันของตัวแปรเดิมตั้งแต่หนึ่งตัวขึ้นไปที่ยกกำลังต่างกันหรือคูณกัน

4.2 สร้างตารางค้นหา สร้างตารางค้นหาหรือเอกสารที่แมปคุณสมบัติพหุนามที่สร้างขึ้นกับตัวแปรเดิมที่สอดคล้องกัน รวมข้อมูลต่างๆ เช่น ชื่อตัวแปร เลขยกกำลัง และการดำเนินการทางคณิตศาสตร์ (เช่น การคูณ การยกกำลัง) ที่ใช้สร้างคุณลักษณะพหุนามแต่ละรายการ ตารางนี้ทำหน้าที่เป็นแนวทางอ้างอิง ช่วยให้สามารถตีความและเข้าใจความหมายเบื้องหลังลักษณะพหุนามแต่ละลักษณะได้ง่าย หากมีการปรับเปลี่ยนกระบวนการสร้างคุณลักษณะพหุนามหรือการเปลี่ยนแปลงชุดข้อมูล ให้อัปเดตตารางการค้นหาตามนั้น ปรับปรุงตารางให้ทันสมัยอยู่เสมอเพื่อให้แน่ใจว่าตารางนั้นสะท้อนถึงความสัมพันธ์ระหว่างตัวแปรเดิมและคุณสมบัติพหุนามได้อย่างถูกต้อง

5. ไฮเปอร์พารามิเตอร์ลิสต์ขั้นตอนกำหนดรายการของไฮเปอร์พารามิเตอร์ที่จะใช้สำหรับอัลกอริธึมเชิงพันธุกรรมและแบบจำลองแรนดอมฟอเรสต์ ตัวอย่าง ได้แก่ n_estimators, max_ความลึก, min_samples_split, max_features, min_samples_leaf และ bootstrap สำหรับอัลกอริทึมทางพันธุกรรม

- Population Size จำนวนบุคคล (โครโมโซม) ในแต่ละรุ่นของอัลกอริทึมทางพันธุกรรม
- Mutation Rate ความน่าจะเป็นที่ยีนในโครโมโซมจะกลายพันธุ์ในระหว่างกระบวนการวิวัฒนาการ
- Crossover Rate ความน่าจะเป็นที่โครโมโซมพ่อแม่ อยู่ระหว่างการผสมข้ามเพื่อผลิตโครโมโซมลูก
- Number of Generations จำนวนรุ่นสูงสุดที่อัลกอริธึมเชิงพันธุกรรมจะวนซ้ำ
- Selection Method วิธีการที่ใช้ในการเลือกโครโมโซมพ่อแม่เพื่อการสืบพันธุ์ เช่น การเลือกการแข่งขันหรือการเลือกวงล้อรูเล็ต

- Fitness Function ฟังก์ชั่นวัตถุประสงค์ที่ใช้ในการประเมินสมรรถภาพหรือประสิทธิภาพของโครโมโซมแต่ละตัว

สำหรับตัวแบบแรนดอมฟอร์เรสต์

-n_estimators จำนวนต้นไม้การตัดสินใจในชุดต้นไม่ตัดสินใจ

-max_depth ความลึกสูงสุด สำหรับแผนผังการตัดสินใจแต่ละรายการในชุด

-min_samples_split จำนวนตัวอย่างขั้นต่ำที่จำเป็นในการดำเนินการแยกที่โหนดในแผนผังการตัดสินใจ

-max_features จำนวนของคุณลักษณะ ที่ต้องพิจารณาเมื่อค้นหาการแยกที่ดีที่สุดในแต่ละโหนดของต้นไม้

- min_samples_leaf จำนวนตัวอย่างขั้นต่ำที่จำเป็นสำหรับโหนดปลายสุดในแผนผังการตัดสินใจ

- bootstrap สุ่มต้นไม้ตัดสินใจบางส่วนของแรนดอมฟอร์เรสต์เพื่อใช้การฝึกสอน

แต่ละไฮเปอร์พารามิเตอร์มีอิทธิพลต่อพฤติกรรมและประสิทธิภาพของอัลกอริทึมทางพันธุกรรมหรือแบบจำลอง แรนดอมฟอร์เรสต์ ในรูปแบบต่างๆ ค่าหรือช่วงเฉพาะสำหรับแต่ละไฮเปอร์พารามิเตอร์สามารถตั้งค่าตามตั้งต้น ความเชี่ยวชาญหรือผ่านกระบวนการค้นหาที่เป็นระบบ เช่น การค้นหาแบบกริดหรือการค้นหาแบบสุ่ม สิ่งสำคัญคือต้องสังเกตว่าตัวเลือกของไฮเปอร์พารามิเตอร์อาจแตกต่างกันไปขึ้นอยู่กับปัญหาเฉพาะ ชุดข้อมูล และอัลกอริทึมหรือตัวแบบที่ใช้ การทดลองและการปรับแต่งไฮเปอร์พารามิเตอร์มักจำเป็นเพื่อเพิ่มประสิทธิภาพการทำงานของอัลกอริทึมทางพันธุกรรมหรือแบบจำลองแรนดอมฟอร์เรสต์ สำหรับงานที่กำหนด

6. การปรับไฮเปอร์พารามิเตอร์และการหาตัวแบบที่เหมาะสม

6.1 การกำหนดปัญหา (เข้ารหัสโครโมโซม)กำหนดรูปแบบปัญหาโดยการเข้ารหัสการแสดงโครโมโซมที่แสดงถึงวิธีแก้ปัญหาที่เป็นไปได้ กำหนดโครงสร้างและองค์ประกอบของโครโมโซม ซึ่งควรจับค่าไฮเปอร์พารามิเตอร์สำหรับอัลกอริทึมทางพันธุกรรมหรือแบบจำลองป่าสุ่ม

6.2 สร้างประชากร (เริ่มต้น) สร้างประชากรเริ่มต้นของโครโมโซมซึ่งเป็นตัวแทนของไฮเปอร์พารามิเตอร์ชุดต่างๆ สำหรับอัลกอริทึมทางพันธุกรรมหรือแบบจำลองป่าสุ่ม ขนาดของกลุ่มประชากรเริ่มต้นขึ้นอยู่กับอัลกอริทึมที่เลือกและความซับซ้อนของปัญหา ตรวจสอบความหลากหลายภายในประชากรเริ่มต้นเพื่อสำรวจชุดค่าผสมไฮเปอร์พารามิเตอร์ที่หลากหลาย

6.3 ประเมินสมรรถภาพ ประเมินความเหมาะสมของโครโมโซมแต่ละตัวโดยการฝึกสอนและประเมินอัลกอริธึมทางพันธุกรรมที่เกี่ยวข้องหรือแบบจำลองแรนดอมฟอร์เรสต์ โดยใช้คุณลักษณะ

ดีกรีของพหุนาม ไฮเปอร์พารามิเตอร์ที่ได้จากการถอดรหัสโครโมโซม เลือกเมตริกการประเมินที่เหมาะสมตามประเภทของปัญหา เช่น ความถูกต้อง หรือ ความแม่นยำ การประเมินความเหมาะสมจะเป็นตัวกำหนดว่าไฮเปอร์พารามิเตอร์ชุดใดชุดหนึ่งทำงานได้ดีเพียงใดสำหรับปัญหาที่กำหนด โดยวิธีการเลือกคุณลักษณะประเภทนี้คือ วิธีแบบแรปเปอร์

6.4 เงื่อนไขการหยุด กำหนดเงื่อนไขการหยุดสำหรับอัลกอริทึมเชิงพันธุกรรม ซึ่งระบุว่าอัลกอริทึมควรยุติเมื่อใด เงื่อนไขการหยุดทำงานทั่วไป ได้แก่ การถึงจำนวนรุ่นสูงสุด การบรรลุเกณฑ์ความเหมาะสมที่ต้องการ หรือเมื่ออัลกอริทึมมาบรรจบกัน เงื่อนไขการหยุดทำให้แน่ใจว่าอัลกอริทึมไม่ได้ทำงานอย่างไม่มีกำหนด และถึงทางออกที่น่าพอใจแล้ว

6.5 ครอสโอเวอร์และการกลายพันธุ์ ใช้ตัวดำเนินการทางพันธุกรรม เช่น ครอสโอเวอร์และการกลายพันธุ์เพื่อสร้างโครโมโซมลูกใหม่จากโครโมโซมแม่ที่เลือก ครอสโอเวอร์เกี่ยวข้องกับการรวมสารพันธุกรรมจากโครโมโซมของผู้ปกครอง 2 โครโมโซมเพื่อสร้างโครโมโซมลูกใหม่ การกลายพันธุ์ทำให้เกิดการเปลี่ยนแปลงแบบสุ่มเล็กน้อยกับโครโมโซมลูกหลานเพื่อสำรวจพื้นที่ใหม่ของพื้นที่แก้ปัญหา

6.6 ผลลัพธ์ หลังจากการวนซ้ำของขั้นตอนวิธีเชิงพันธุกรรมเสร็จสิ้นหรือตรงตามเงื่อนไขการหยุด ให้แสดงผลลัพธ์ออกมา ระบุโครโมโซมที่มีค่าความเหมาะสมสูงสุด ซึ่งแสดงถึงชุดของไฮเปอร์พารามิเตอร์ที่ปรับให้เหมาะสมที่สุด บันทึกโครโมโซมที่มีประสิทธิภาพดีที่สุดและค่าความเหมาะสมที่สอดคล้องกันเพื่อใช้อ้างอิงในอนาคต ชุดพารามิเตอร์ไฮเปอร์พารามิเตอร์ที่ปรับให้เหมาะสมนี้สามารถใช้เพื่อฝึกอัลกอริทึมทางพันธุกรรมขั้นสุดท้ายหรือแบบจำลองแรนดอมฟอร์เรสต์อัลกอริธึมเชิงพันธุกรรมและกระบวนการปรับแต่งไฮเปอร์พารามิเตอร์มีจุดประสงค์เพื่อค้นหาชุดค่าผสมที่ดีที่สุดของไฮเปอร์พารามิเตอร์ที่เพิ่มประสิทธิภาพสูงสุดของอัลกอริทึมเชิงพันธุกรรมหรือตัวแบบแรนดอมฟอร์เรสต์สำหรับปัญหาที่กำหนด อัลกอริทึมจะค้นหาโซลูชันที่เหมาะสมที่สุดภายในพื้นที่โซลูชันที่กำหนดด้วยการพัฒนาและประเมินประชากรของโครโมโซมซ้ำๆ

สิ่งสำคัญคือต้องสังเกตว่าลักษณะเฉพาะของการใช้อัลกอริทึมเชิงพันธุกรรม เช่น วิธีการเลือกตัวดำเนินการแบบไขว้และการกลายพันธุ์ และการตั้งค่าพารามิเตอร์ อาจแตกต่างกันไปขึ้นอยู่กับปัญหาและไลบรารีหรือเฟรมเวิร์กของอัลกอริธึมเชิงพันธุกรรมที่ใช้ ในทำนองเดียวกัน กระบวนการปรับแต่งไฮเปอร์พารามิเตอร์สามารถดำเนินการได้หลายวิธี รวมถึงการค้นหากริด การค้นหาแบบสุ่มหรือเทคนิคขั้นสูงอื่นๆ เช่น การปรับให้เหมาะสมแบบเบย์

7. ตารางผลลัพธ์

ตารางผลลัพธ์ให้ข้อมูลสรุปและระเบียบเกี่ยวกับประสิทธิภาพของตัวแบบ และช่วยให้เปรียบเทียบและวิเคราะห์ได้ง่าย ช่วยในการระบุรูปแบบต่างๆ ที่มีประสิทธิภาพสูงสุดหรือชุดค่าผสมของไฮเปอร์พารามิเตอร์ และให้ข้อมูลเชิงลึกที่มีค่าสำหรับการวิเคราะห์หรือการตัดสินใจเพิ่มเติม

สำหรับการดำเนินการในขั้นต่อไป จะเป็นการทดสอบเฟรมเวิร์ค ที่ได้นำเสนอกับชุดข้อมูลที่ได้นำเสนอ โดยผลการทดลองทั้งหมด จะอยู่ในบทที่ 4

3.2.3 การประยุกต์ใช้ขนาดของล็อกไฟล์เพื่อการเลือกไฮเปอร์พารามิเตอร์และแบบจำลองที่เหมาะสมที่สุด

สืบเนื่องจากเฟรมเวิร์คและขั้นตอนวิธีการเลือกคุณสมบัติ ดีกรี และไฮเปอร์พารามิเตอร์ที่นำเสนอ มุ่งเน้นไปที่วัตถุประสงค์ในการสร้างตารางผลลัพธ์ความถูกต้องที่ได้จากการลดลงของจำนวนคุณลักษณะเพื่อให้ผู้ใช้เลือกแบบจำลองที่ตรงกับความต้องการ ซึ่งมักจะไม่เกิดความคลุมเครือในกรณีที่ผู้ใช้มีวัตถุประสงค์อย่างใดอย่างหนึ่งที่ชัดเจน แต่ในกรณีที่ผู้ใช้เกิดความไม่ชัดเจนในการกำหนดวัตถุประสงค์ระหว่างความถูกต้อง จำนวนคุณลักษณะ หรือแม้แต่ขนาดของล็อกไฟล์ อาจส่งผลให้การเลือกแบบจำลองของหลายผู้ใช้เกิดความขัดแย้งกันเอง

เพื่อให้มีมาตรฐานในการเลือกแบบจำลอง วิทยานิพนธ์จึงขอเสนอแนะแนวทางในการพิจารณาค่าความถูกต้องและขนาดของล็อกไฟล์ในปัญหาการหาค่าเหมาะสุดแบบหลายวัตถุประสงค์ โดยมี 2 วัตถุประสงค์หลักคือ ค่าความถูกต้องและขนาดของล็อกไฟล์
แบบจำลองทางคณิตศาสตร์สำหรับปัญหาดังกล่าวเป็นดังนี้

กำหนดให้

s_{max} คือขนาดล็อกไฟล์ที่ใหญ่ที่สุดในอุดมคติ

a_{max} คือค่าความถูกต้องที่สูงที่สุดในอุดมคติ

s_i คือขนาดของล็อกไฟล์ที่ได้จากค่าไฮเปอร์พารามิเตอร์ที่พิจารณา

a_i คือค่าความถูกต้องที่ได้จากค่าไฮเปอร์พารามิเตอร์ที่พิจารณา

โดยในวิทยานิพนธ์นี้ขอกำหนดค่า s_{max} ให้เป็นค่าขนาดของล็อกไฟล์สะอาดและ a_{max} ให้เป็นค่า 1.0 ที่แสดงถึงความเป็นไปได้สูงสุดของค่าความถูกต้อง
วัตถุประสงค์ของแบบจำลองแบ่งเป็นสองกรอบความคิดคือ

1. ต้องการลดขนาดของล็อกไฟล์ให้ได้มากที่สุด หรือ $s_{max} - s_i$ มีค่ามากที่สุด
2. ต้องการลดความถูกต้องให้น้อยที่สุด หรือ $a_{max} - a_i$ มีค่าน้อยที่สุด

ทั้งสองวัตถุประสงค์สามารถถูกพิจารณาร่วมกันด้วยการกำหนดฟังก์ชันวัตถุประสงค์ดังนี้

$$\boldsymbol{Maximize}\ \frac{s_{max}-s_i}{a_{max}-a_i} \qquad (9)$$

เมื่อพิจารณาที่ฟังก์ชันวัตถุประสงค์จะเห็นว่า เมื่อ $s_{max} - s_i$ มีค่ามากขึ้นฟังก์ชันวัตถุประสงค์ก็จะมีค่ามากตาม และเมื่อ $a_{max} - a_i$ มีค่าน้อยฟังก์ชันวัตถุประสงค์จะผกผันมีค่า

มากขึ้น จึงสามารถใช้ฟังก์ชันวัตถุประสงค์เป็นตัวแทนของการแก้ปัญหาการหาค่าเหมาะสุดแบบหลายวัตถุประสงค์ได้

บทที่ 4

ผลการวิจัย

ในบทนี้จะขอแสดงผลการทดลองโดยเริ่มจากการวิเคราะห์คุณลักษณะผ่านการให้คะแนนด้วยวิธีไคสแควร์ การวิเคราะห์ความแปรปรวน และสารสนเทศร่วม แล้วจึงแสดงผลลัพธ์ที่ได้จากการใช้เฟรมเวิร์คการเลือกคุณลักษณะสองขั้นตอน ผลลัพธ์จากขั้นตอนวิธีเชิงพันธุกรรม โดยผลลัพธ์ที่ได้จากเฟรมเวิร์คจะถูกนำมาพิจารณาร่วมกับขนาดของล็อกไฟล์เพื่อหาการจัดเก็บล็อกไฟล์ที่เหมาะสุด ซึ่งจะถูกนำเสนอเป็นส่วนสุดท้าย

4.1 การวิเคราะห์คุณลักษณะ

เพื่อวัตถุประสงค์ของการจัดเก็บล็อกไฟล์ ในวิทยาพนธ์นี้จะขอวิเคราะห์คุณลักษณะสำหรับสองชุดข้อมูล คือ ชุดข้อมูลไอโอที และชุดข้อมูลเอ็นเอสแอลเคดีดี โดยตาราง 22 ถึงตาราง 24 แสดงค่าคะแนนของคุณลักษณะด้วยวิธีไคสแควร์ การวิเคราะห์ความแปรปรวน และสารสนเทศร่วมสำหรับชุดข้อมูลไอโอที ตามลำดับ และตาราง 25 ถึงตาราง 27 แสดงค่าคะแนนของคุณลักษณะด้วยสามวิธีเดียวกันสำหรับชุดข้อมูลเอ็นเอสแอลเคดีดี นอกจากนี้ตารางแสดงค่าคะแนนที่ได้จากวิธีไคสแควร์และการวิเคราะห์ความแปรปรวนที่เป็นวิธีทางสถิติจะมีการแสดงค่าความน่าจะเป็นประกอบด้วย ซึ่งสามารถนำมาวิเคราะห์นัยสำคัญทางสถิติได้ด้วยการตั้งสมมติฐานว่าง เพื่อทดสอบว่าตัวแปรต้นหรือคุณลักษณะเป็นอิสระต่อตัวแปรตามหรือค่าเป้าหมายหรือไม่ ในอีกความหมายคือทั้งคุณลักษณะและค่าเป้าหมายไม่มีความสัมพันธ์กัน

การตรวจสอบสมมติฐานว่างทำได้ด้วยการตั้งค่าระดับความมั่นใจ ยกตัวอย่างเช่น ระดับความเชื่อมั่นที่ 95% ค่าความน่าจะเป็นที่จะพิสูจน์สมมติฐานจะต้องเป็นค่าที่น้อยกว่า 0.05 กล่าวคือหากคุณลักษณะใดมีค่าความน่าจะเป็นที่น้อยกว่า 0.05 จะถือว่ามีความสัมพันธ์กับค่าเป้าหมาย แต่สำหรับคุณลักษณะที่มีค่าความน่าจะเป็นมากกว่าหรือเท่ากับ 0.05 จะถือว่าไม่มีความสัมพันธ์กับค่าเป้าหมาย

จากการวิเคราะห์ค่าความน่าจะเป็นพบว่าทั้งสองชุดข้อมูลประกอบด้วยคุณลักษณะที่มีความสัมพันธ์กับค่าเป้าหมายแบบมีนัยยะ โดยค่าความน่าจะเป็นที่ได้จากวิธีไคสแควร์และการวิเคราะห์ความแปรปรวนสำหรับชุดข้อมูลไอโอทีที่มีค่าน้อยกว่า 0.05 มีจำนวน 67 และ 62 สำหรับชุดข้อมูลเอ็นเอสแอลเคดีดีมีจำนวน 105 และ 102 ตามลำดับ (แสดงโดยเส้นคู่แบ่งแถว) ซึ่งจำนวนดังกล่าวควรจะถูกนำมาพิจารณาในการกำหนดจำนวนคุณลักษณะ กล่าวคือจำนวนคุณลักษณะที่จะ

เลือกไม่ควรมีค่าเกินกว่าค่าจำนวนที่หามาได้ ดังนั้นการเลือกจำนวนคุณลักษณะในวิทยานิพนธ์นี้ที่ 20 และ 40 จึงเป็นคุณลักษณะที่มีนัยสำคัญ

นอกจากนี้ยังจะขอแสดงค่าคะแนนและค่าความน่าจะเป็นในลักษณะแผนภูมิเส้นใน

ภาพ **6** ถึง

ภาพ **15** เพื่อให้เห็นความเปลี่ยนแปลงของค่าดังกล่าว จุดที่น่าสังเกตคือลักษณะของการหักศอกที่สามารถบ่งบอกถึงคะแนนของคุณลักษณะที่ลดลงอย่างหนักถือเป็นจุดเปลี่ยนที่อาจนำมาประกอบการเลือกจำนวนของคุณลักษณะ

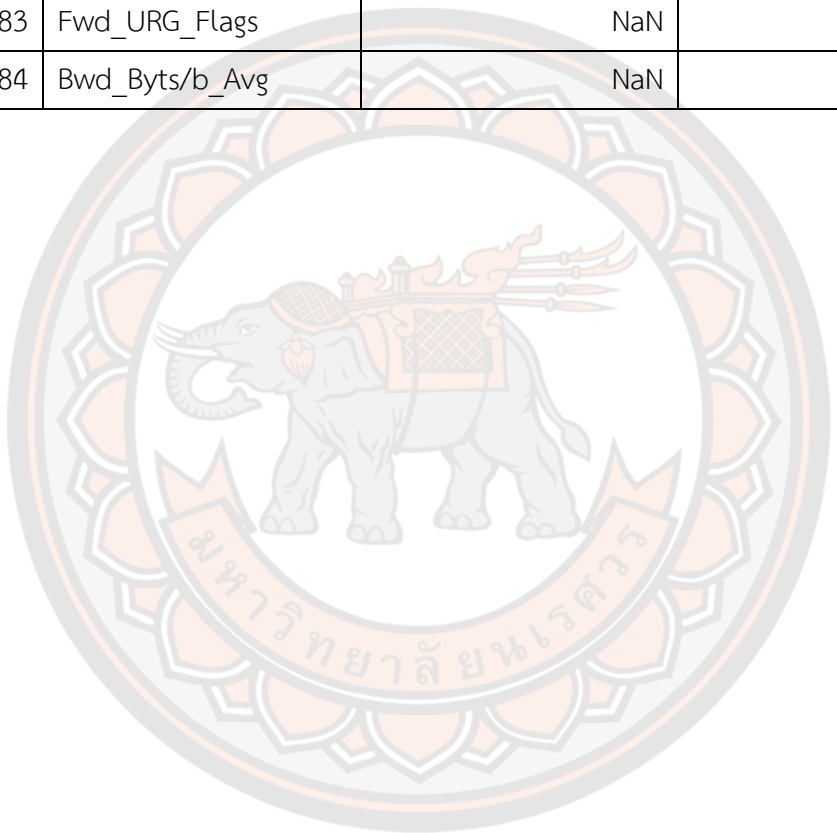
ตาราง 22 ค่าคะแนนของคุณลักษณะด้วยวิธีไคสแควร์ของชุดข้อมูลไอโอที

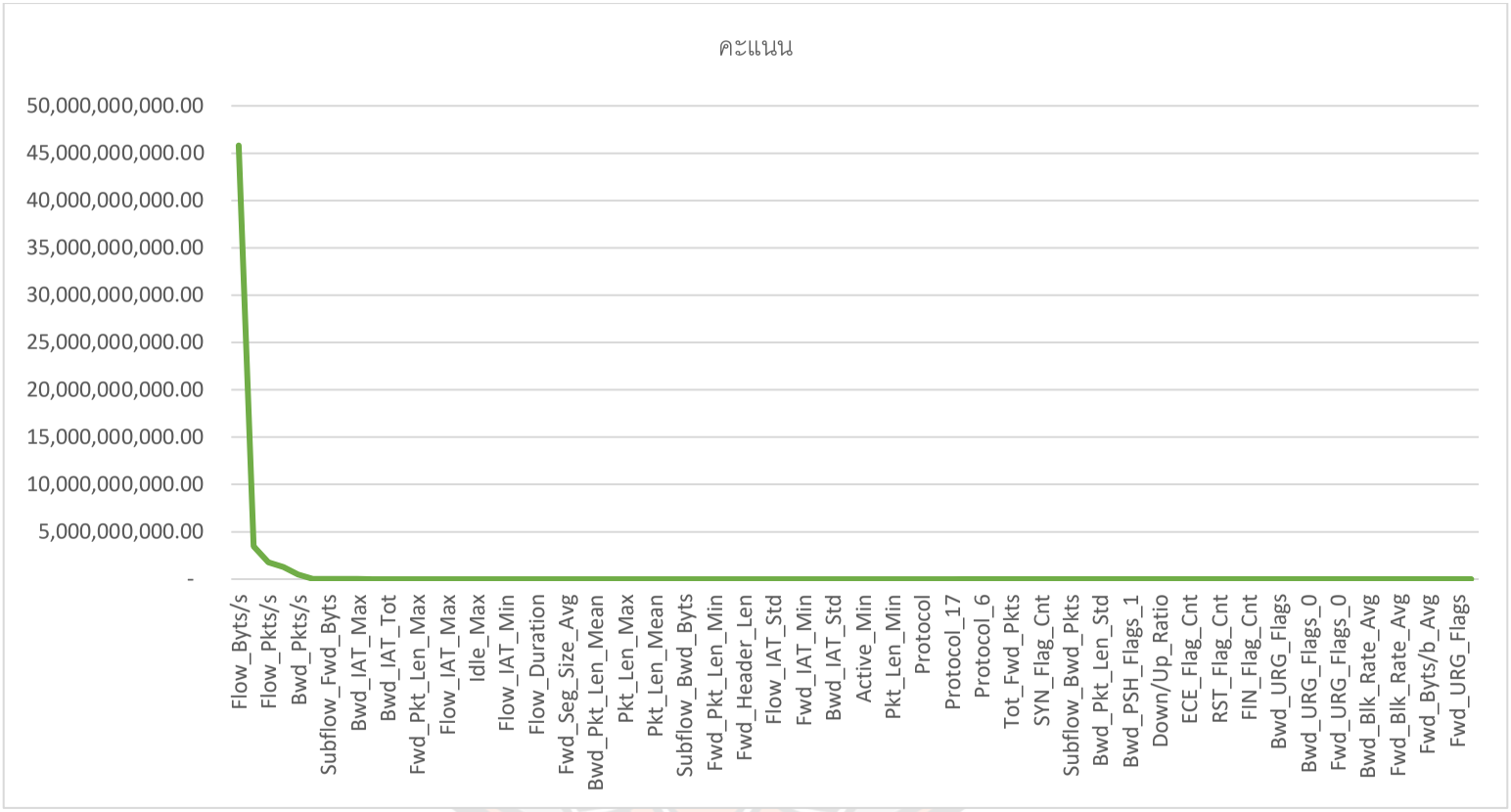
ลำดับ	คุณลักษณะ	คะแนน	ความน่าจะเป็น
1	Flow_Byts/s	45,817,327,611.84	0.000
2	Pkt_Len_Var	3,452,929,147.41	0.000
3	Flow_Pkts/s	1,782,429,682.03	0.000
4	Fwd_Pkts/s	1,290,780,303.01	0.000
5	Bwd_Pkts/s	502,996,033.57	0.000
6	Fwd_Pkt_Len_Std	25,818,667.78	0.000
7	Subflow_Fwd_Byts	19,347,882.44	0.000
8	TotLen_Fwd_Pkts	19,347,882.44	0.000
9	Bwd_IAT_Max	8,780,776.14	0.000
10	Bwd_IAT_Min	8,468,215.16	0.000
11	Bwd_IAT_Tot	8,449,576.01	0.000
12	Bwd_IAT_Mean	8,417,603.25	0.000
13	Fwd_Pkt_Len_Max	8,061,757.05	0.000
14	Pkt_Len_Std	5,352,236.10	0.000
15	Flow_IAT_Max	5,178,534.54	0.000
16	Idle_Min	5,161,873.15	0.000
17	Idle_Max	5,093,710.85	0.000
18	Idle_Mean	4,958,562.89	0.000

ลำดับ	คุณลักษณะ	คะแนน	ความน่าจะเป็น
19	Flow_IAT_Min	4,537,786.74	0.000
20	Flow_IAT_Mean	4,481,867.36	0.000
21	Flow_Duration	4,140,481.92	0.000
22	Fwd_Pkt_Len_Mean	4,117,514.62	0.000
23	Fwd_Seg_Size_Avg	4,117,514.62	0.000
24	Bwd_Pkt_Len_Min	3,009,886.67	0.000
25	Bwd_Pkt_Len_Mean	2,880,330.78	0.000
26	Bwd_Seg_Size_Avg	2,880,330.78	0.000
27	Pkt_Len_Max	2,792,672.31	0.000
28	Bwd_Pkt_Len_Max	2,729,423.66	0.000
29	Pkt_Len_Mean	1,647,978.38	0.000
30	TotLen_Bwd_Pkts	1,587,279.59	0.000
31	Subflow_Bwd_Byts	1,587,279.59	0.000
32	Pkt_Size_Avg	1,473,505.85	0.000
33	Fwd_Pkt_Len_Min	494,362.10	0.000
34	Bwd_Header_Len	340,580.34	0.000
35	Fwd_Header_Len	337,030.25	0.000
36	Fwd_IAT_Std	324,913.04	0.000
37	Flow_IAT_Std	197,093.55	0.000
38	Fwd_IAT_Max	163,636.40	0.000
39	Fwd_IAT_Min	131,389.18	0.000
40	Active_Max	128,810.34	0.000
41	Bwd_IAT_Std	128,308.12	0.000
42	Active_Mean	114,081.42	0.000
43	Active_Min	104,897.32	0.000
44	Fwd_IAT_Tot	96,725.22	0.000
45	Pkt_Len_Min	65,817.87	0.000
46	Idle_Std	50,797.21	0.000
47	Protocol	46,905.49	0.000

ลำดับ	คุณลักษณะ	คะแนน	ความน่าจะเป็น
48	ACK_Flag_Cnt	12,303.51	0.000
49	Protocol_17	10,859.42	0.000
50	Active_Std	10,581.11	0.000
51	Protocol_6	6,704.75	0.000
52	Subflow_Fwd_Pkts	3,958.84	0.000
53	Tot_Fwd_Pkts	3,958.84	0.000
54	Fwd_Act_Data_Pkts	3,099.02	0.000
55	SYN_Flag_Cnt	2,783.11	0.000
56	Fwd_IAT_Mean	2,494.42	0.000
57	Subflow_Bwd_Pkts	905.759	0.000
58	Tot_Bwd_Pkts	905.759	0.000
59	Bwd_Pkt_Len_Std	715.661	0.000
60	Bwd_PSH_Flags	599.029	0.000
61	Bwd_PSH_Flags_1	599.029	0.000
62	PSH_Flag_Cnt	599.029	0.000
63	Down/Up_Ratio	336.247	0.000
64	Protocol_0	265.687	0.000
65	ECE_Flag_Cnt	67.601	0.000
66	Bwd_PSH_Flags_0	16.333	0.000
67	RST_Flag_Cnt	8.388	0.004
68	CWE_Flag_Count	1.975	0.160
69	FIN_Flag_Cnt	1.126	0.289
70	URG_Flag_Cnt	0.133	0.716
71	Bwd_URG_Flags	0.133	0.716
72	Bwd_URG_Flags_1	0.133	0.716
73	Bwd_URG_Flags_0	0.000	0.998
74	Fwd_PSH_Flags_0	0.000	1.000
75	Fwd_URG_Flags_0	0.000	1.000
76	Fwd_Seg_Size_Min	NaN	NaN

ลำดับ	คุณลักษณะ	คะแนน	ความน่าจะเป็น
77	Bwd_Blk_Rate_Avg	NaN	NaN
78	Bwd_Pkts/b_Avg	NaN	NaN
79	Fwd_Blk_Rate_Avg	NaN	NaN
80	Fwd_Pkts/b_Avg	NaN	NaN
81	Fwd_Byts/b_Avg	NaN	NaN
82	Fwd_PSH_Flags	NaN	NaN
83	Fwd_URG_Flags	NaN	NaN
84	Bwd_Byts/b_Avg	NaN	NaN





ภาพ 6 แผนภูมิเส้นแสดงคะแนนของทุกคุณลักษณะด้วยวิธีไคสแควร์ของชุดข้อมูลไอโอที

ภาพ 7 แผนภูมิเส้นแสดงความน่าจะเป็นของทุกคุณลักษณะวิธีโคลแควร์ของชุดข้อมูลไอโอที

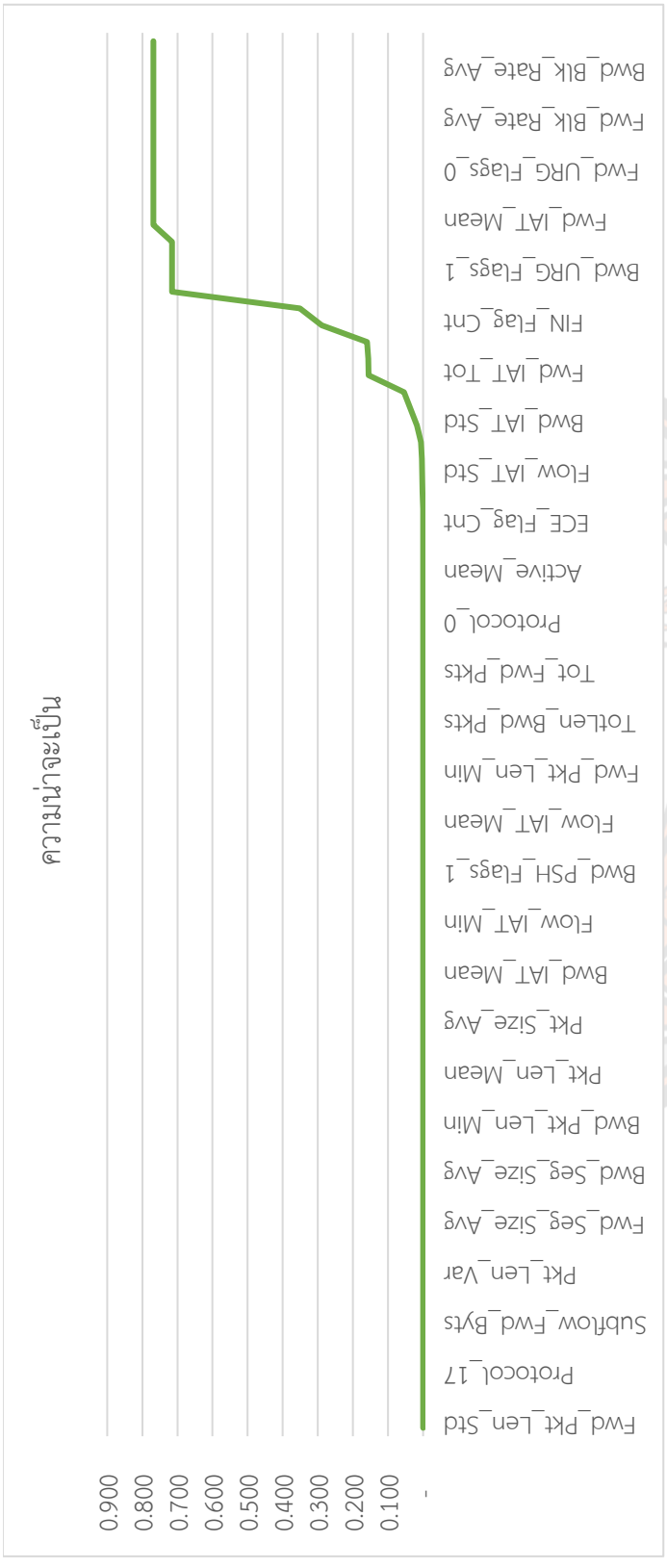
ตาราง 23 ค่าคะแนนของคุณลักษณะด้วยวิธีการวิเคราะห์ความแปรปรวนของชุดข้อมูลไอโอที

ลำดับ	คุณลักษณะ	คะแนน	ความน่าจะเป็น
1	Fwd_Pkt_Len_Std	37,898.192	0.000
2	ACK_Flag_Cnt	27,895.072	0.000
3	Protocol_6	18,585.799	0.000
4	Protocol_17	17,826.564	0.000
5	Protocol	16,783.414	0.000
6	Pkt_Len_Std	9,479.644	0.000
7	Subflow_Fwd_Byts	8,603.524	0.000
8	TotLen_Fwd_Pkts	8,603.524	0.000
9	Fwd_Pkt_Len_Max	8,406.182	0.000
10	Pkt_Len_Var	7,627.821	0.000
11	Bwd_Header_Len	5,819.120	0.000
12	Fwd_Header_Len	4,468.281	0.000
13	Fwd_Seg_Size_Avg	4,367.205	0.000
14	Fwd_Pkt_Len_Mean	4,367.205	0.000
15	Bwd_Pkt_Len_Mean	4,136.541	0.000
16	Bwd_Seg_Size_Avg	4,136.541	0.000
17	Pkt_Len_Max	4,068.310	0.000
18	Bwd_Pkt_Len_Max	3,889.344	0.000
19	Bwd_Pkt_Len_Min	3,834.539	0.000
20	SYN_Flag_Cnt	3,090.643	0.000
21	Flow_Pkts/s	2,874.615	0.000
22	Pkt_Len_Mean	2,470.808	0.000
23	Fwd_Pkts/s	2,380.404	0.000
24	Bwd_Pkts/s	2,296.047	0.000
25	Pkt_Size_Avg	1,505.201	0.000
26	Flow_Byts/s	1,427.166	0.000
27	Bwd_IAT_Min	774.734	0.000

ลำดับ	คุณลักษณะ	คะแนน	ความน่าจะเป็น
28	Bwd_IAT_Mean	756.331	0.000
29	Subflow_Bwd_Pkts	714.829	0.000
30	Tot_Bwd_Pkts	714.829	0.000
31	Flow_IAT_Min	672.082	0.000
32	Idle_Min	638.430	0.000
33	Bwd_PSH_Flags	616.226	0.000
34	Bwd_PSH_Flags_1	616.226	0.000
35	PSH_Flag_Cnt	616.226	0.000
36	Bwd_PSH_Flags_0	616.226	0.000
37	Flow_IAT_Mean	600.753	0.000
38	Bwd_IAT_Max	572.070	0.000
39	Idle_Mean	554.754	0.000
40	Fwd_Pkt_Len_Min	498.464	0.000
41	Down/Up_Ratio	493.206	0.000
42	Bwd_IAT_Tot	447.919	0.000
43	TotLen_Bwd_Pkts	399.278	0.000
44	Subflow_Bwd_Byts	399.278	0.000
45	Subflow_Fwd_Pkts	358.850	0.000
46	Tot_Fwd_Pkts	358.850	0.000
47	Flow_IAT_Max	358.691	0.000
48	Idle_Max	350.079	0.000
49	Protocol_0	268.543	0.000
50	Fwd_Act_Data_Pkts	250.654	0.000
51	Flow_Duration	217.394	0.000
52	Active_Mean	92.668	0.000
53	Active_Min	89.785	0.000
54	Pkt_Len_Min	78.807	0.000
55	ECE_Flag_Cnt	67.615	0.000
56	Active_Max	63.654	0.000

ลำดับ	คุณลักษณะ	คะแนน	ความน่าจะเป็น
57	Fwd_IAT_Std	9.760	0.002
58	Flow_IAT_Std	9.263	0.002
59	RST_Flag_Cnt	8.392	0.004
60	Active_Std	7.576	0.006
61	Bwd_IAT_Std	5.604	0.018
62	Fwd_IAT_Min	4.387	0.036
63	Fwd_IAT_Max	3.713	0.054
64	Fwd_IAT_Tot	2.015	0.156
65	Idle_Std	2.012	0.156
66	CWE_Flag_Count	1.975	0.160
67	FIN_Flag_Cnt	1.126	0.289
68	Bwd_Pkt_Len_Std	0.869	0.351
69	Bwd_URG_Flags_0	0.133	0.716
70	Bwd_URG_Flags_1	0.133	0.716
71	Bwd_URG_Flags	0.133	0.716
72	URG_Flag_Cnt	0.133	0.716
73	Fwd_IAT_Mean	0.087	0.769
74	Fwd_PSH_Flags	NaN	NaN
75	Fwd_URG_Flags	NaN	NaN
76	Fwd_URG_Flags_0	NaN	NaN
77	Fwd_Byts/b_Avg	NaN	NaN
78	Fwd_Pkts/b_Avg	NaN	NaN
79	Fwd_Blk_Rate_Avg	NaN	NaN
80	Bwd_Byts/b_Avg	NaN	NaN
81	Bwd_Pkts/b_Avg	NaN	NaN
82	Bwd_Blk_Rate_Avg	NaN	NaN
83	Fwd_Seg_Size_Min	NaN	NaN
84	Fwd_PSH_Flags_0	NaN	NaN

ภาพ 8 แผนภูมิเส้นแสดงคะแนนของทุกคุณลักษณะด้วยวิธีการวิเคราะห์ความแปรปรวนของชุดข้อมูลไอโอที



ภาพ 9 แผนภูมิเส้นแสดงความน่าจะเป็นของทุกคุณลักษณะด้วยวิธีการวิเคราะห์ความแปรปรวนของชุดข้อมูลไอโอที

ตาราง 24 ค่าคะแนนของคุณลักษณะด้วยวิธีสารสนเทศร่วมของชุดข้อมูลไอโอที

ลำดับ	คุณลักษณะ	คะแนน
1	Flow_Duration	0.136
2	Flow_Pkts/s	0.121
3	Idle_Mean	0.114
4	Flow_IAT_Mean	0.112
5	Flow_Byts/s	0.111
6	Bwd_Pkts/s	0.108
7	Idle_Max	0.103
8	Flow_IAT_Max	0.101
9	TotLen_Bwd_Pkts	0.097
10	Subflow_Bwd_Byts	0.097
11	Fwd_PSH_Flags_0	0.096
12	Fwd_URG_Flags_0	0.096
13	Bwd_URG_Flags_0	0.096
14	TotLen_Fwd_Pkts	0.093
15	Subflow_Fwd_Byts	0.093
16	Fwd_Seg_Size_Avg	0.092
17	Fwd_Pkt_Len_Mean	0.092
18	Pkt_Size_Avg	0.092
19	Bwd_PSH_Flags_0	0.091
20	Fwd_Pkt_Len_Max	0.090
21	Pkt_Len_Mean	0.085
22	Fwd_Pkt_Len_Min	0.084
23	Protocol	0.078
24	Bwd_Pkt_Len_Mean	0.075
25	Bwd_Seg_Size_Avg	0.075
26	Bwd_Header_Len	0.068
27	Fwd_Pkts/s	0.067

ลำดับ	คุณลักษณะ	คะแนน
28	Bwd_Pkt_Len_Min	0.067
29	Idle_Min	0.066
30	ACK_Flag_Cnt	0.065
31	Protocol_6	0.065
32	Pkt_Len_Max	0.063
33	Bwd_Pkt_Len_Max	0.061
34	Flow_IAT_Min	0.059
35	Pkt_Len_Min	0.057
36	Bwd_IAT_Tot	0.057
37	Fwd_Header_Len	0.053
38	Pkt_Len_Var	0.050
39	Bwd_IAT_Mean	0.049
40	Pkt_Len_Std	0.049
41	Bwd_IAT_Max	0.049
42	Subflow_Bwd_Pkts	0.045
43	Tot_Bwd_Pkts	0.045
44	Protocol_17	0.043
45	Idle_Std	0.040
46	Flow_IAT_Std	0.039
47	Fwd_Pkt_Len_Std	0.038
48	Fwd_IAT_Tot	0.034
49	Fwd_IAT_Mean	0.034
50	Bwd_IAT_Min	0.031
51	Fwd_IAT_Max	0.028
52	Fwd_IAT_Min	0.020
53	Bwd_IAT_Std	0.019
54	Subflow_Fwd_Pkts	0.016
55	Tot_Fwd_Pkts	0.016
56	Fwd_IAT_Std	0.014

ลำดับ	คุณลักษณะ	คะแนน
57	Bwd_Pkt_Len_Std	0.014
58	Down/Up_Ratio	0.014
59	Fwd_Act_Data_Pkts	0.013
60	SYN_Flag_Cnt	0.008
61	Active_Max	0.003
62	Active_Mean	0.002
63	Active_Min	0.002
64	Bwd_PSH_Flags_1	0.001
65	Active_Std	0.001
66	Protocol_0	0.001
67	Bwd_PSH_Flags	0.001
68	Fwd_Seg_Size_Min	0.001
69	Fwd_URG_Flags	0.001
70	RST_Flag_Cnt	0.001
71	Bwd_URG_Flags_1	0.001
72	FIN_Flag_Cnt	0.000
73	Bwd_Byts/b_Avg	0.000
74	PSH_Flag_Cnt	0.000
75	Fwd_PSH_Flags	0.000
76	Fwd_Blk_Rate_Avg	0.000
77	Bwd_Blk_Rate_Avg	0.000
78	Bwd_Pkts/b_Avg	0.000
79	Bwd_URG_Flags	0.000
80	Fwd_Pkts/b_Avg	0.000
81	Fwd_Byts/b_Avg	0.000
82	ECE_Flag_Cnt	0.000
83	CWE_Flag_Count	0.000
84	URG_Flag_Cnt	0.000

ภาพ 10 แผนภูมิเส้นแสดงคะแนนของทุกคุณลักษณะด้วยวิธีสารสนเทศร่วมของชุดข้อมูลไอโอที

ตาราง 25 ค่าคะแนนของคุณลักษณะด้วยวิธีไคสแควร์ของชุดข้อมูลเอ็นเอสแอลเคดีดี

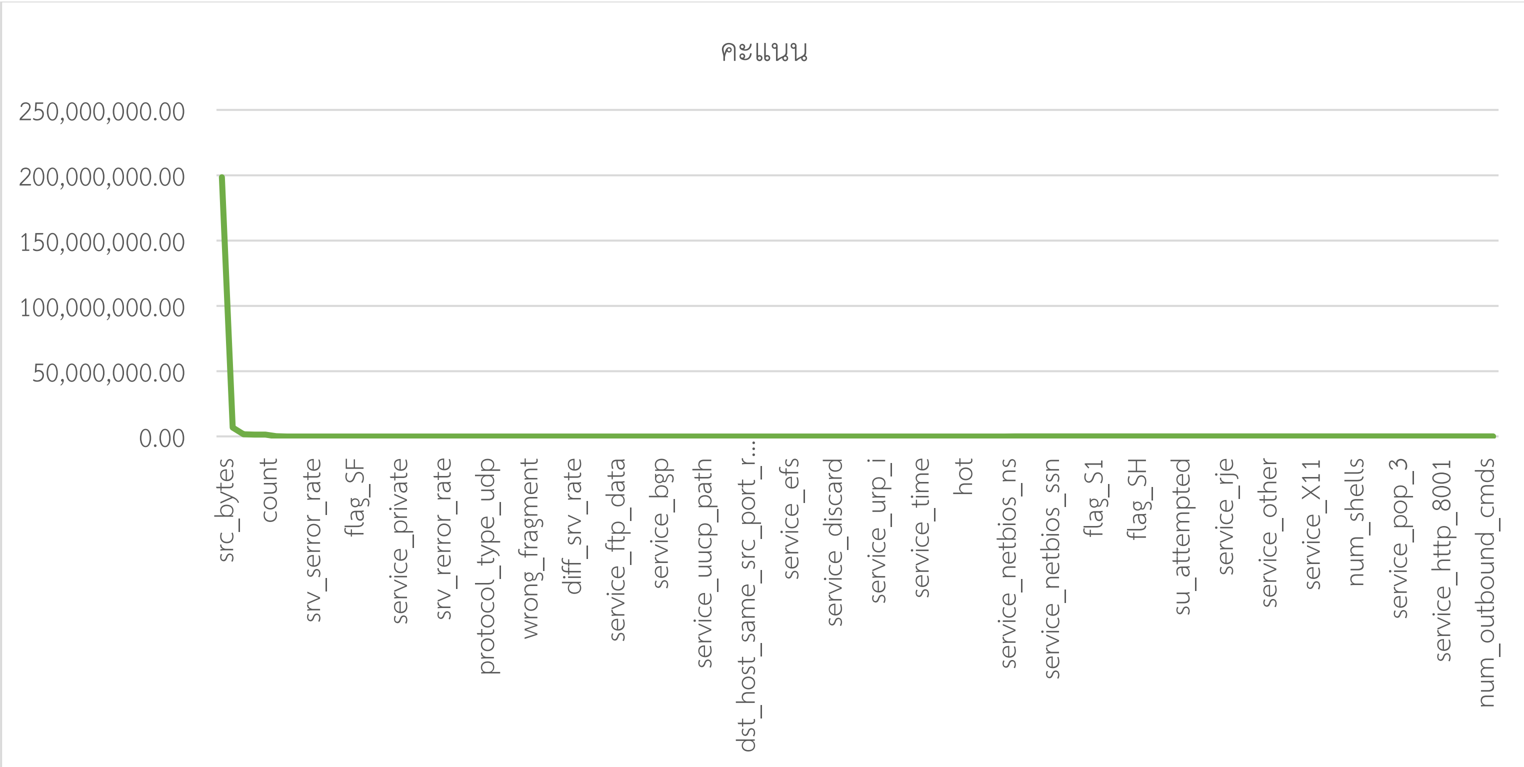
ลำดับ	คุณลักษณะ	คะแนน	ความน่าจะเป็น
1	src_bytes	198,442,495.29	0.000
2	dst_bytes	6,824,753.65	0.000
3	duration	1,544,251.90	0.000
4	dst_host_srv_count	1,386,748.79	0.000
5	count	1,311,861.11	0.000
6	dst_host_count	183,980.16	0.000
7	flag_S0	7,687.68	0.000
8	dst_host_srv_serror_rate	7,655.57	0.000
9	srv_serror_rate	7,464.04	0.000
10	serror_rate	7,436.20	0.000
11	dst_host_serror_rate	7,408.82	0.000
12	logged_in	7,218.84	0.000
13	flag_SF	5,851.82	0.000
14	service_http	5,439.96	0.000
15	num_root	5,203.64	0.000
16	dst_host_same_srv_rate	4,680.36	0.000
17	service_private	4,165.86	0.000
18	num_compromised	4,159.92	0.000
19	same_srv_rate	4,137.73	0.000
20	service_domain_u	1,577.91	0.000
21	srv_rerror_rate	1,428.01	0.000
22	rerror_rate	1,423.39	0.000
23	dst_host_srv_rerror_rate	1,410.70	0.000
24	dst_host_rerror_rate	1,314.34	0.000
25	protocol_type_udp	1,079.93	0.000
26	service_smtp	1,053.82	0.000
27	protocol_type_icmp	941.001	0.000

ลำดับ	คุณลักษณะ	คะแนน	ความน่าจะเป็น
28	flag_REJ	809.251	0.000
29	wrong_fragment	684.876	0.000
30	service_eco_i	680.271	0.000
31	dst_host_diff_srv_rate	606.642	0.000
32	service_ecr_i	548.480	0.000
33	diff_srv_rate	482.311	0.000
34	flag_RSTR	455.407	0.000
35	srv_diff_host_rate	251.647	0.000
36	service_Z39_50	196.988	0.000
37	service_ftp_data	188.252	0.000
38	service_courier	187.826	0.000
39	flag_RSTO	184.971	0.000
40	service_uucp	179.809	0.000
41	service_bgp	167.211	0.000
42	service_whois	166.065	0.000
43	num_file_creations	161.050	0.000
44	service_imap4	158.048	0.000
45	service_uucp_path	152.322	0.000
46	service_iso_tsap	150.031	0.000
47	service_ctf	145.450	0.000
48	service_nnsp	140.869	0.000
49	dst_host_same_src_port_rate	140.427	0.000
50	service_supdup	130.562	0.000
51	service_http_443	129.416	0.000
52	service_csnet_ns	127.126	0.000
53	service_efs	125.981	0.000
54	service_gopher	124.835	0.000
55	service_vmnet	122.545	0.000
56	service_daytime	122.545	0.000

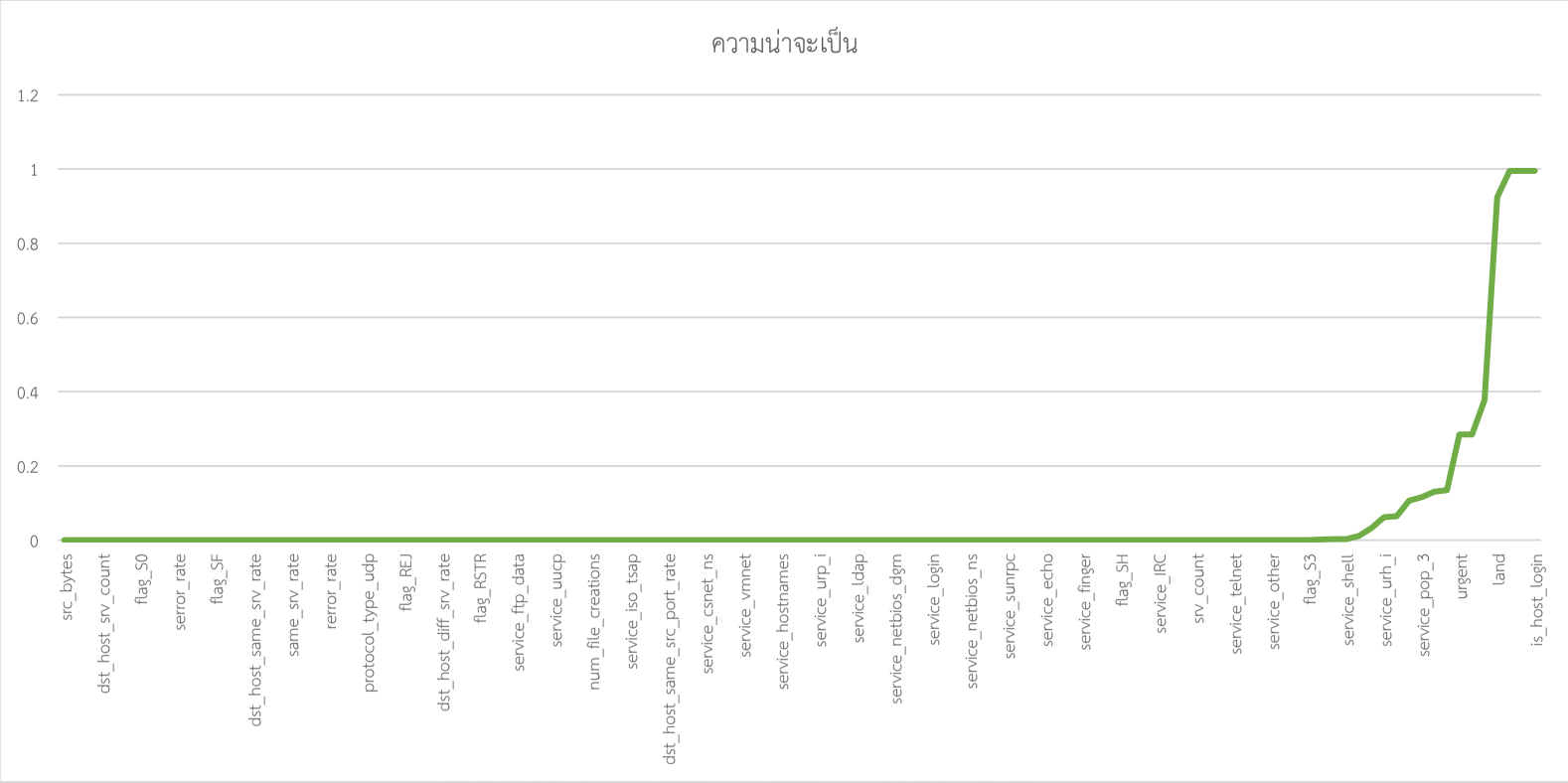
ลำดับ	คุณลักษณะ	คะแนน	ความน่าจะเป็น
57	service_discard	120.254	0.000
58	service_hostnames	109.947	0.000
59	service_klogin	105.366	0.000
60	service_name	105.366	0.000
61	service_urp_i	104.557	0.000
62	service_exec	104.220	0.000
63	service_mtp	103.075	0.000
64	service_ldap	103.075	0.000
65	service_time	102.077	0.000
66	service_systat	100.784	0.000
67	service_netbios_dgm	97.349	0.000
68	service_link	97.349	0.000
69	hot	97.305	0.000
70	service_login	90.477	0.000
71	service_netstat	89.332	0.000
72	service_domain	89.206	0.000
73	service_netbios_ns	87.041	0.000
74	num_access_files	77.364	0.000
75	service_kshell	76.734	0.000
76	service_sunrpc	76.734	0.000
77	service_netbios_ssn	76.734	0.000
78	service_auth	76.285	0.000
79	service_echo	74.443	0.000
80	service_nntp	69.862	0.000
81	flag_S1	69.527	0.000
82	service_finger	67.473	0.000
83	service_ssh	66.426	0.000
84	service_sql_net	52.683	0.000
85	flag_SH	49.247	0.000

ลำดับ	คุณลักษณะ	คะแนน	ความน่าจะเป็น
86	dst_host_srv_diff_host_rate	38.301	0.000
87	is_guest_login	37.311	0.000
88	service_IRC	34.926	0.000
89	su_attempted	29.687	0.000
90	service_ntp_u	27.941	0.000
91	srv_count	26.832	0.000
92	flag_RSTOS0	24.051	0.000
93	service_rje	22.906	0.000
94	service_telnet	19.857	0.000
95	service_pop_2	19.470	0.000
96	service_remote_job	19.470	0.000
97	service_other	17.398	0.000
98	protocol_type_tcp	15.022	0.000
99	service_printer	13.743	0.000
100	flag_S3	13.097	0.000
101	service_X11	9.614	0.002
102	root_shell	8.682	0.003
103	service_shell	8.673	0.003
104	flag_S2	6.413	0.011
105	num_shells	4.559	0.033
106	service_urh_i	3.493	0.062
107	service_pm_dump	3.436	0.064
108	service_red_i	2.619	0.106
109	service_pop_3	2.468	0.116
110	service_tim_i	2.291	0.130
111	flag_OTH	2.240	0.135
112	urgent	1.145	0.285
113	service_http_8001	1.145	0.285
114	service_ftp	0.780	0.377

ลำดับ	คุณลักษณะ	คะแนน	ความน่าจะเป็น
115	land	0.009	0.924
116	num_failed_logins	0.000	0.995
117	num_outbound_cmds	0.000	0.995
118	is_host_login	0.000	0.995



ภาพ 11 แผนภูมิเส้นแสดงคะแนนของทุกคุณลักษณะด้วยวิธีไคสแควร์ ของชุดข้อมูลเอ็นเอสแอลเคดีดี



ภาพ 12 แผนภูมิเส้นแสดงความน่าจะเป็นของทุกคุณลักษณะวิธีไคสแควร์ของชุดข้อมูลเอ็นเอสแอลเคดีดี

ตาราง 26 ค่าคะแนนของคุณลักษณะด้วยวิธีการวิเคราะห์ความแปรปรวน ของชุดข้อมูลเอ็นเอสแอลเคดีดี

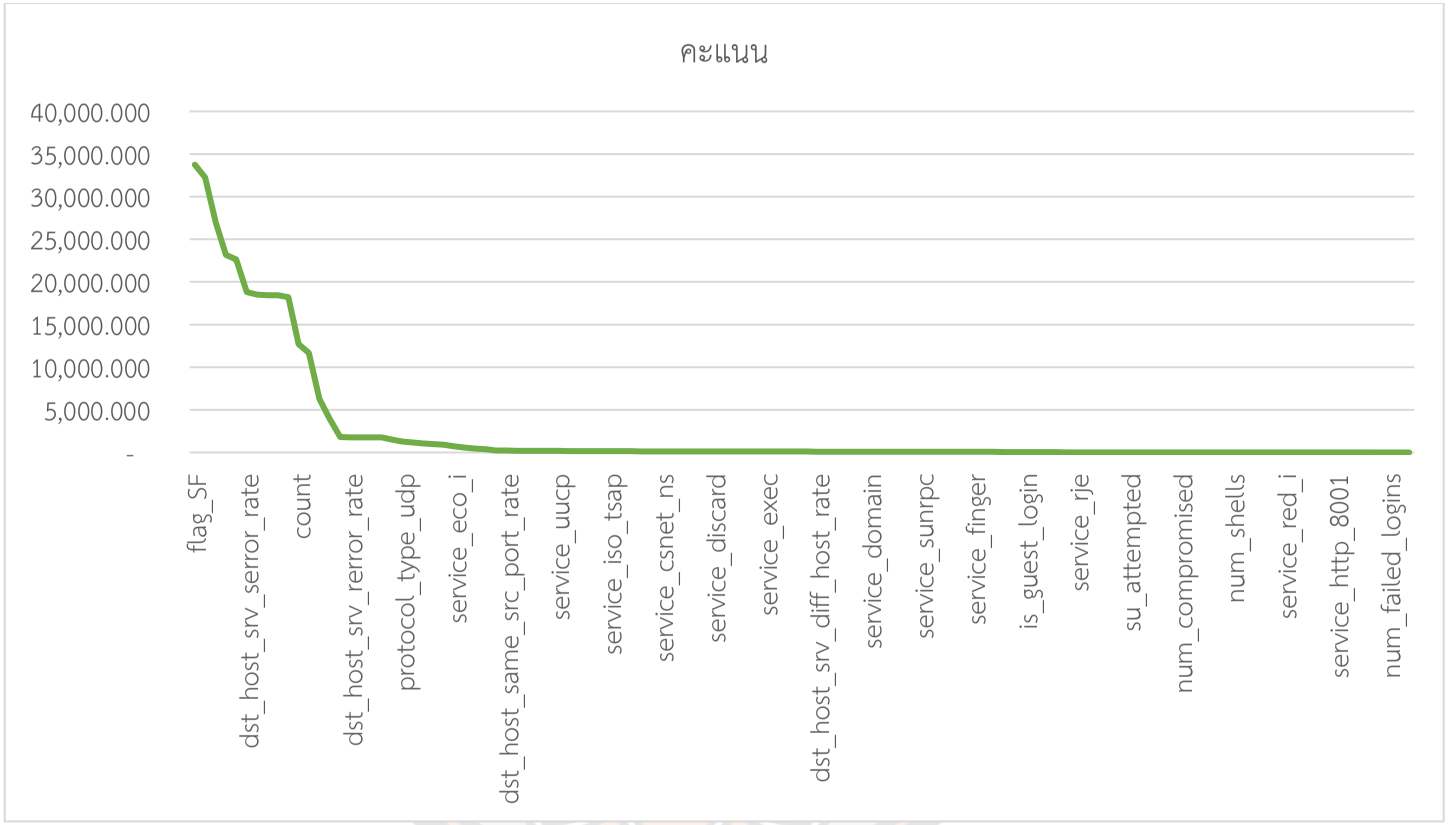
ลำดับ	คุณลักษณะ	คะแนน	ความน่าจะเป็น
1	flag_SF	33,753.437	0.000
2	same_srv_rate	32,237.024	0.000
3	dst_host_srv_count	27,004.333	0.000
4	dst_host_same_srv_rate	23,173.921	0.000
5	logged_in	22,650.570	0.000
6	dst_host_srv_serror_rate	18,802.401	0.000
7	dst_host_serror_rate	18,532.466	0.000
8	flag_S0	18,451.328	0.000
9	serror_rate	18,424.350	0.000
10	srv_serror_rate	18,216.145	0.000
11	count	12,689.566	0.000
12	service_http	11,663.317	0.000
13	service_private	6,293.078	0.000
14	dst_host_count	3,966.245	0.000
15	service_domain_u	1,823.783	0.000
16	dst_host_srv_rerror_rate	1,781.421	0.000
17	rerror_rate	1,779.331	0.000
18	dst_host_rerror_rate	1,771.239	0.000
19	srv_rerror_rate	1,768.869	0.000
20	dst_host_diff_srv_rate	1,514.826	0.000
21	protocol_type_udp	1,289.192	0.000
22	service_smtp	1,169.969	0.000
23	protocol_type_icmp	1,049.028	0.000
24	diff_srv_rate	980.157	0.000
25	flag_REJ	919.622	0.000
26	service_eco_i	726.019	0.000

ลำดับ	คุณลักษณะ	คะแนน	ความน่าจะเป็น
27	service_ecr_i	574.945	0.000
28	flag_RSTR	473.263	0.000
29	srv_diff_host_rate	372.089	0.000
30	wrong_fragment	242.386	0.000
31	dst_host_same_src_port_rate	219.645	0.000
32	service_ftp_data	200.869	0.000
33	service_Z39_50	199.900	0.000
34	service_courier	190.471	0.000
35	flag_RSTO	188.617	0.000
36	service_uucp	182.231	0.000
37	service_bgp	169.302	0.000
38	service_whois	168.128	0.000
39	service_imap4	159.915	0.000
40	service_uucp_path	154.055	0.000
41	service_iso_tsap	151.712	0.000
42	service_ctf	147.029	0.000
43	service_nnsp	142.349	0.000
44	service_supdup	131.831	0.000
45	service_http_443	130.663	0.000
46	service_csnet_ns	128.329	0.000
47	service_efs	127.162	0.000
48	service_gopher	125.995	0.000
49	service_vmnet	123.662	0.000
50	service_daytime	123.662	0.000
51	service_discard	121.330	0.000
52	service_hostnames	110.844	0.000
53	service_name	106.189	0.000
54	service_klogin	106.189	0.000
55	service_urp_i	105.506	0.000

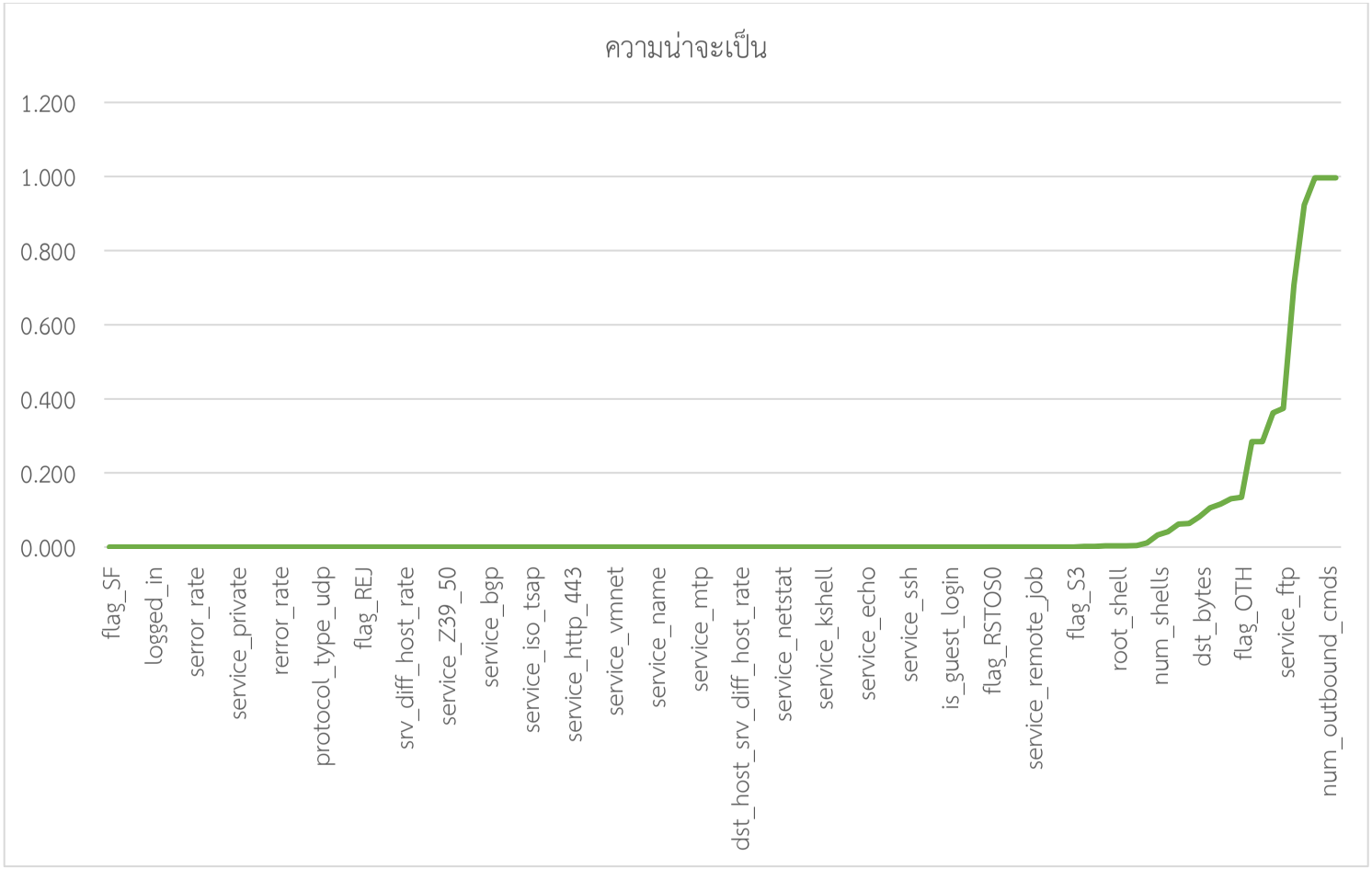
ลำดับ	คุณลักษณะ	คะแนน	ความน่าจะเป็น
56	service_exec	105.026	0.000
57	service_mtp	103.863	0.000
58	service_ldap	103.863	0.000
59	service_time	103.121	0.000
60	service_systat	101.537	0.000
61	dst_host_srv_diff_host_rate	100.147	0.000
62	service_netbios_dgm	98.051	0.000
63	service_link	98.051	0.000
64	service_login	91.083	0.000
65	service_netstat	89.922	0.000
66	service_domain	89.907	0.000
67	service_netbios_ns	87.601	0.000
68	protocol_type_tcp	81.363	0.000
69	service_kshell	77.168	0.000
70	service_netbios_ssn	77.168	0.000
71	service_sunrpc	77.168	0.000
72	service_auth	77.091	0.000
73	service_echo	74.851	0.000
74	service_nntp	70.221	0.000
75	flag_S1	69.959	0.000
76	service_finger	68.648	0.000
77	service_ssh	66.751	0.000
78	duration	65.435	0.000
79	service_sql_net	52.886	0.000
80	flag_SH	49.424	0.000
81	is_guest_login	37.709	0.000
82	service_IRC	35.027	0.000
83	num_access_files	34.530	0.000
84	service_ntp_u	28.005	0.000

ลำดับ	คุณลักษณะ	คะแนน	ความน่าจะเป็น
85	flag_RSTOS0	24.092	0.000
86	service_rje	22.943	0.000
87	service_telnet	20.260	0.000
88	service_pop_2	19.496	0.000
89	service_remote_job	19.496	0.000
90	service_other	18.023	0.000
91	su_attempted	16.845	0.000
92	service_printer	13.756	0.000
93	flag_S3	13.111	0.000
94	num_root	9.832	0.002
95	service_X11	9.626	0.002
96	num_compromised	8.737	0.003
97	root_shell	8.698	0.003
98	service_shell	8.679	0.003
99	num_file_creations	8.459	0.004
100	flag_S2	6.419	0.011
101	num_shells	4.561	0.033
102	hot	4.153	0.042
103	service_urh_i	3.493	0.062
104	service_pm_dump	3.436	0.064
105	dst_bytes	3.020	0.082
106	service_red_i	2.620	0.106
107	service_pop_3	2.473	0.116
108	service_tim_i	2.291	0.130
109	flag_OTH	2.240	0.134
110	urgent	1.145	0.285
111	service_http_8001	1.145	0.285
112	src_bytes	0.831	0.362
113	service_ftp	0.790	0.374

ลำดับ	คุณลักษณะ	คะแนน	ความน่าจะเป็น
114	srv_count	0.142	0.707
115	land	0.009	0.924
116	num_failed_logins	0.000	0.996
117	num_outbound_cmds	0.000	0.996
118	is_host_login	0.000	0.996



ภาพ 13 แผนภูมิเส้นแสดงคะแนนของทุกคุณลักษณะด้วยวิธีการวิเคราะห์ความแปรปรวนของชุดข้อมูลเอ็นเอสแอลเคดีดี



ภาพ 14 แผนภูมิเส้นแสดงความน่าจะเป็นของทุกคุณลักษณะด้วยวิธีการวิเคราะห์ความแปรปรวนของชุดข้อมูลเอ็นเอสแอลเคดีดี

ตาราง 27 ค่าคะแนนของคุณลักษณะด้วยวิธีสารสนเทศร่วมของชุดข้อมูลเอ็นเอสแอลเคดีดี

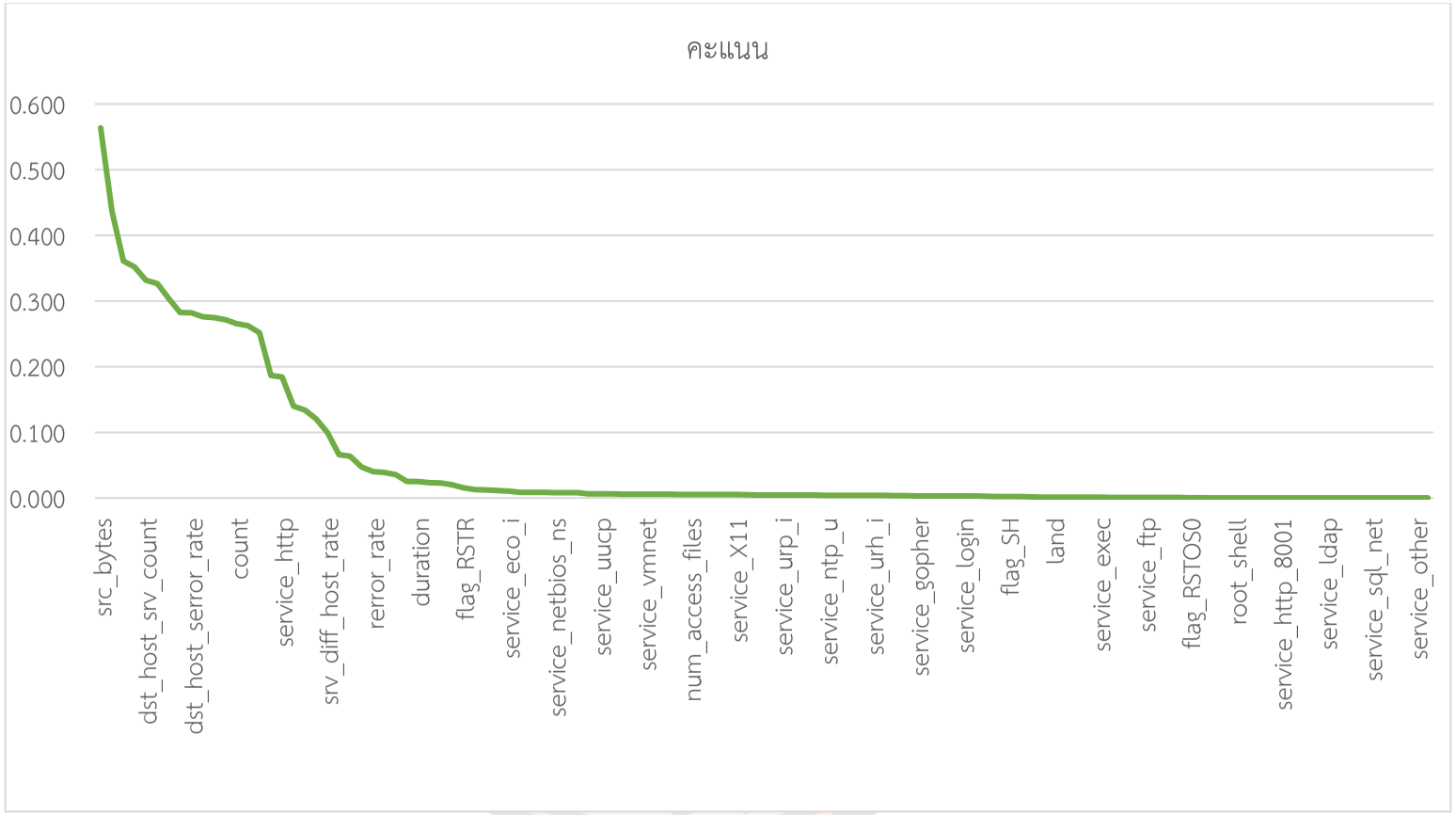
ลำดับ	คุณลักษณะ	คะแนน
1	src_bytes	0.564
2	dst_bytes	0.437
3	diff_srv_rate	0.361
4	same_srv_rate	0.352
5	dst_host_srv_count	0.331
6	flag_SF	0.326
7	dst_host_same_srv_rate	0.304
8	dst_host_diff_srv_rate	0.283
9	dst_host_serror_rate	0.282
10	dst_host_srv_serror_rate	0.276
11	logged_in	0.274
12	serror_rate	0.271
13	count	0.265
14	srv_serror_rate	0.262
15	flag_S0	0.252
16	dst_host_srv_diff_host_rate	0.187
17	service_http	0.184
18	dst_host_count	0.140
19	dst_host_same_src_port_rate	0.134
20	service_private	0.120
21	srv_diff_host_rate	0.099
22	srv_count	0.066
23	dst_host_srv_rerror_rate	0.063
24	service_domain_u	0.047
25	rerror_rate	0.040
26	dst_host_rerror_rate	0.039
27	srv_rerror_rate	0.036

ลำดับ	คุณลักษณะ	คะแนน
28	service_smtp	0.025
29	duration	0.025
30	protocol_type_udp	0.023
31	protocol_type_icmp	0.023
32	flag_REJ	0.020
33	flag_RSTR	0.016
34	service_ecr_i	0.013
35	hot	0.012
36	service_efs	0.011
37	service_eco_i	0.010
38	service_auth	0.009
39	num_compromised	0.008
40	service_imap4	0.008
41	service_netbios_ns	0.008
42	service_ctf	0.008
43	service_ftp_data	0.008
44	num_root	0.006
45	service_uucp	0.006
46	service_Z39_50	0.006
47	flag_S2	0.006
48	protocol_type_tcp	0.006
49	service_vmnet	0.006
50	service_domain	0.006
51	service_iso_tsap	0.006
52	service_link	0.005
53	num_access_files	0.005
54	service_nntp	0.005
55	service_netstat	0.005
56	service_finger	0.005

ลำดับ	คุณลักษณะ	คะแนน
57	service_X11	0.005
58	service_supdup	0.005
59	service_netbios_dgm	0.004
60	service_hostnames	0.004
61	service_urp_i	0.004
62	num_outbound_cmds	0.004
63	service_echo	0.004
64	service_http_443	0.004
65	service_ntp_u	0.004
66	service_nnsp	0.004
67	flag_OTH	0.004
68	service_red_i	0.004
69	service_urh_i	0.004
70	service_bgp	0.004
71	service_kshell	0.003
72	service_pop_3	0.003
73	service_gopher	0.003
74	service_shell	0.003
75	num_shells	0.003
76	service_whois	0.003
77	service_login	0.003
78	service_csnet_ns	0.003
79	service_daytime	0.002
80	service_netbios_ssn	0.002
81	flag_SH	0.002
82	service_name	0.002
83	service_rje	0.002
84	flag_RSTO	0.001
85	land	0.001

ลำดับ	คุณลักษณะ	คะแนน
86	wrong_fragment	0.001
87	service_IRC	0.001
88	service_discard	0.001
89	service_exec	0.001
90	flag_S1	0.001
91	service_sunrpc	0.001
92	flag_S3	0.001
93	service_ftp	0.001
94	service_telnet	0.001
95	service_time	0.001
96	service_klogin	0.001
97	flag_RSTOS0	0.000
98	service_systat	0.000
99	urgent	0.000
100	num_file_creations	0.000
101	root_shell	0.000
102	su_attempted	0.000
103	num_failed_logins	0.000
104	service_pm_dump	0.000
105	service_http_8001	0.000
106	service_uucp_path	0.000
107	is_host_login	0.000
108	service_pop_2	0.000
109	service_ldap	0.000
110	service_tim_i	0.000
111	service_mtp	0.000
112	service_ssh	0.000
113	service_sql_net	0.000
114	service_courier	0.000

ลำดับ	คุณลักษณะ	คะแนน
115	service_remote_job	0.000
116	service_printer	0.000
117	service_other	0.000
118	is_guest_login	0.000



ภาพ 15 แผนภูมิเส้นแสดงคะแนนของทุกคุณลักษณะด้วยวิธีสารสนเทศร่วมของชุดข้อมูลไอโอที

4.3 ผลการทดลองของเฟรมเวิร์คการเลือกคุณลักษณะสองขั้นตอน

ผลลัพธ์ที่ได้จากการดำเนินกระบวนการทั้งหมดตั้งแต่ M1 ถึง M4 จะถูกรวบรวมเพื่อนำมาพิจารณาเลือกค่าไฮเปอร์พารามิเตอร์ที่เหมาะสม โดยในส่วนนี้ความต้องการของผู้ใช้จะมีกรอบความคิดสองกรอบ กล่าวคือความต้องการความถูกต้องที่สูงที่สุด และความต้องการใช้จำนวนคุณลักษณะให้น้อยที่สุด สำหรับกรณีที่สองผลลัพธ์ที่ได้จะถูกเลือกจากกรณีที่จำนวนคุณลักษณะ 20 จำนวนแล้วแข่งกันด้วยค่าความถูกต้องที่สูงที่สุด โดยค่าที่มีความเกี่ยวข้องกับการเป็นคำตอบจะถูกแสดงด้วยตัวอักษรหนาในตารางผลลัพธ์

ตาราง 28 แสดงตารางผลลัพธ์ของชุดข้อมูลไอโอที โดยกรอบความคิดแรกจะได้การทำกระบวนการ M1 โดยที่ไม่มีการคัดกรองคุณลักษณะเป็นคำตอบ ในขณะที่กรอบความคิดที่สองจะได้กระบวนการ M2 ที่ผ่านการเลือกคุณลักษณะ 20 ตัวด้วยการวิเคราะห์ความแปรปรวนแล้วจึงทำการค้นหาไฮเปอร์พารามิเตอร์เป็นคำตอบ

ตาราง 29 แสดงตารางผลลัพธ์ของชุดข้อมูลเอ็นเอสแอลเคดีดี โดยทั้งสองกรอบความคิดได้คำตอบเดียวกันคือการทำกระบวนการ M4 ที่ผ่านการเลือกคุณลักษณะ 20 ตัวด้วยสารสนเทศร่วมแล้วจึงนำมาสร้างคุณลักษณะพหุนามดีกรี 3 ต่อด้วยการเลือกคุณลักษณะพหุนาม 20 ตัวเพื่อทำการค้นหาไฮเปอร์พารามิเตอร์

สำหรับสองชุดข้อมูลที่เหลือจะถูกนำมาทำการทดลองเพื่อแสดงความทั่วไปของเฟรมเวิร์ครวมถึงข้อจำกัดที่อาจเกิดขึ้นจากความหลากหลายของชุดข้อมูล

ตาราง 30 แสดงตารางผลลัพธ์ของชุดข้อมูลมะเร็ง สำหรับกรอบความคิดแรกจะได้การทำกระบวนการ M3 โดยใช้ความสำคัญของคุณลักษณะในการเลือกคุณลักษณะ 30 ตัวแล้วจึงนำมาสร้างคุณลักษณะพหุนามดีกรี 3 เพื่อทำการค้นหาไฮเปอร์พารามิเตอร์ สำหรับกรอบความคิดที่สองจะได้กระบวนการ M4 เป็นคำตอบ โดยใช้การวิเคราะห์ความแปรปรวนในการเลือกคุณสมบัติ แล้วจึงนำมาสร้างคุณสมบัติพหุนามดีกรี 3 ต่อด้วยการเลือกคุณลักษณะพหุนามก่อนการค้นหาไฮเปอร์พารามิเตอร์

ตาราง31 แสดงตารางผลลัพธ์ของชุดข้อมูลลายมือที่มีลักษณะพิเศษคือ เป็นข้อมูลค่าสีในจุดสีซึ่งเป็นค่าประเภทเดียวกันทั้งหมด สำหรับกรอบความคิดแรกจะได้การทำกระบวนการ M1 เป็นคำตอบ ในขณะที่กระบวนการ M2 เป็นคำตอบสำหรับกรอบความคิดที่สอง โดยใช้ความสำคัญของคุณลักษณะแล้วจึงทำการค้นหาไฮเปอร์พารามิเตอร์

จากผลการเลือกคำตอบข้างต้นจะเห็นได้ว่าเฟรมเวิร์คการเลือกคุณลักษณะสองขั้นตอนที่นำเสนอมีความทั่วไปอันจะสามารถจัดหาข้อมูลที่เป็นประโยชน์กับผู้ใช้ที่จะสามารถเลือกคำตอบได้เหมาะสมกับความต้องการที่ต่างกันสังเกตได้จากการที่คำตอบนั้นกระจายครบไปสู่กระบวนการ M1, M2, M3 และ M4

ตาราง 28 ผลการทดลองของชุดข้อมูลไอโอที

วิธีการเลือกคุณลักษณะ	จำนวนคุณลักษณะ	ดีกรี	M1	M2	M3	M4
ไม่ใช้	**84**	-	**0.953**			
ไคสแควร์	20	2		0.898	0.899	0.866
	20	3		0.899	0.902	0.879
	40	2		0.924	0.913	0.913
	40	3		0.922	0.909	0.909
การวิเคราะห์ความแปรปรวน	20	2		0.947	0.945	0.927
	20	3		**0.948**	0.942	0.921
	40	2		0.938	0.930	0.917
	40	3		0.936	0.926	0.902
สารสนเทศร่วม	20	2		0.919	0.929	0.914
	20	3		0.923	0.922	0.904
	40	2		0.936	0.930	0.924
	40	3		0.934	0.928	0.902
ความสำคัญของคุณลักษณะ	20	2		0.927	0.929	0.912
	20	3		0.926	0.923	0.892
	40	2		0.934	0.935	0.924
	40	3		0.934	0.930	0.902

ตาราง 29 ผลการทดลองของชุดข้อมูลเอ็นเอสแอลเคดีดี

วิธีการเลือกคุณลักษณะ	จำนวนคุณลักษณะ	ดีกรี	M1	M2	M3	M4
ไม่ใช้	118	-	0.775			
ไคสแควร์	20	2		0.785	0.787	0.787
	20	3		0.785	0.785	0.802
	40	2		0.777	0.782	0.772
	40	3		0.778	0.780	0.774
การวิเคราะห์ความแปรปรวน	20	2		0.731	0.737	0.718
	20	3		0.772	0.778	0.793
	40	2		0.758	0.762	0.747
	40	3		0.757	0.757	0.721
สารสนเทศร่วม	20	2		0.778	0.783	0.772
	20	**3**		0.779	0.781	**0.805**
	40	2		0.777	0.786	0.771
	40	3		0.777	0.787	0.775
ความสำคัญของคุณลักษณะ	20	2		0.772	0.778	0.773
	20	3		0.772	0.778	0.804
	40	2		0.778	0.786	0.765
	40	3		0.778	0.786	0.773

ตาราง 30 ผลการทดลองของชุดข้อมูลมะเร็ง

วิธีการเลือกคุณลักษณะ	จำนวนคุณลักษณะ	ดีกรี	M1	M2	M3	M4
ไม่ใช้	30	-	0.964			
ไคสแควร์	20	2		0.965	0.961	0.945
	20	3		0.966	0.960	0.960
	30	2		0.962	0.965	0.951
	30	3		0.961	0.975	0.971
การวิเคราะห์ความแปรปรวน	20	2		0.964	0.965	0.952
	20	**3**		0.964	0.968	**0.970**
	30	2		0.962	0.971	0.952
	30	3		0.962	0.976	0.975
สารสนเทศร่วม	20	2		0.956	0.964	0.953
	20	3		0.957	0.967	0.968
	30	2		0.961	0.971	0.951
	30	3		0.962	0.974	0.973
ความสำคัญของคุณลักษณะ	20	2		0.961	0.961	0.952
	20	3		0.961	0.967	0.966
	30	2		0.961	0.971	0.949
	30	**3**		0.962	**0.977**	0.971

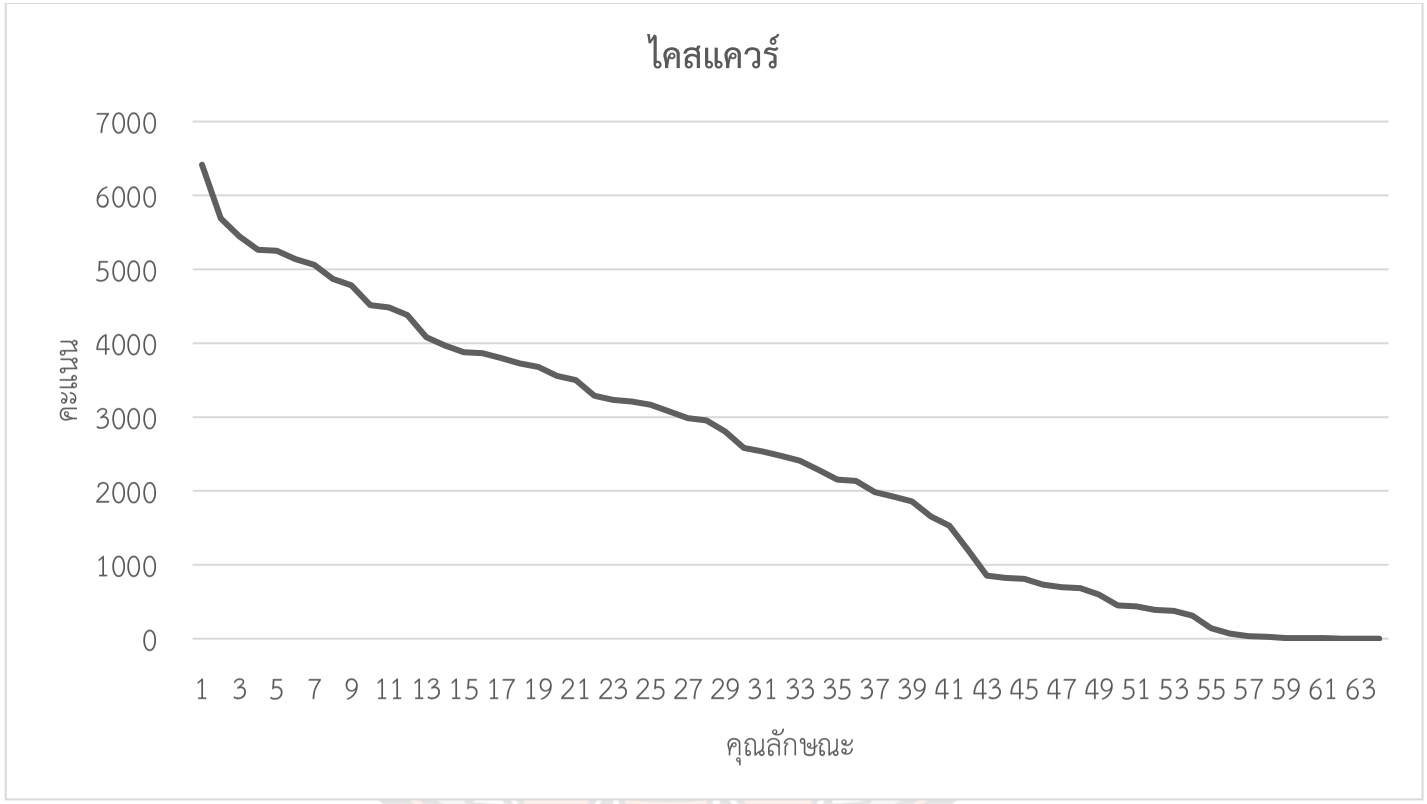
ตาราง 31 ผลการทดลองของชุดข้อมูลลายมือ

วิธีการเลือกคุณลักษณะ	จำนวนคุณลักษณะ	ดีกรี	M1	M2	M3	M4
ไม่ใช้	**118**	-	**0.981**			
ไคสแควร์	20	2		0.956	0.943	0.868
	20	3		0.957	0.935	0.845
	40	2		0.978	0.974	0.932
	40	3		0.978	0.967	0.920
การวิเคราะห์ความแปรปรวน	20	2		0.945	0.934	0.879
	20	3		0.946	0.929	0.864
	40	2		0.978	0.975	0.931
	40	3		0.978	0.971	0.920
สารสนเทศร่วม	20	2		0.952	0.940	0.886
	20	3		0.953	0.933	0.864
	40	2		0.978	0.974	0.935
	40	3		0.978	0.971	0.914
ความสำคัญของคุณลักษณะ	**20**	2		**0.962**	0.952	0.896
	20	3		0.961	0.947	0.871
	40	2		0.979	0.975	0.931
	40	3		0.979	0.970	0.917

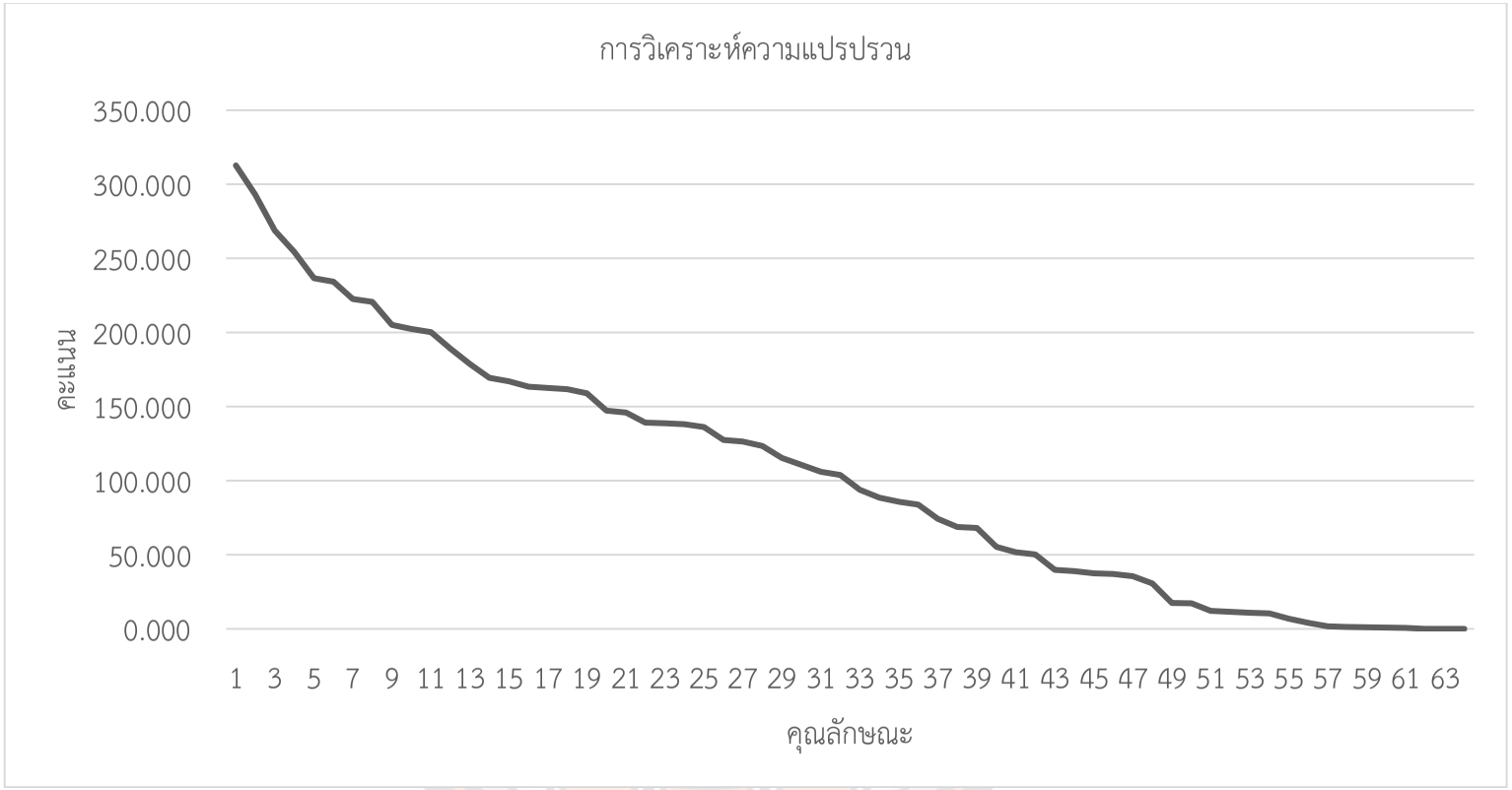
เนื่องด้วยข้อมูลลายมือเป็นข้อมูลที่ทุกคุณลักษณะเป็นประเภทเดียวกันทั้งหมด เพื่อวิเคราะห์ข้อจำกัดของเฟรมเวิร์คที่นำเสนอ จึงทำการวิเคราะห์เพิ่มเติมในส่วนของคะแนนของคุณลักษณะดังแสดงด้วยแผนภูมิเส้นใน

ภาพ **16** ถึง

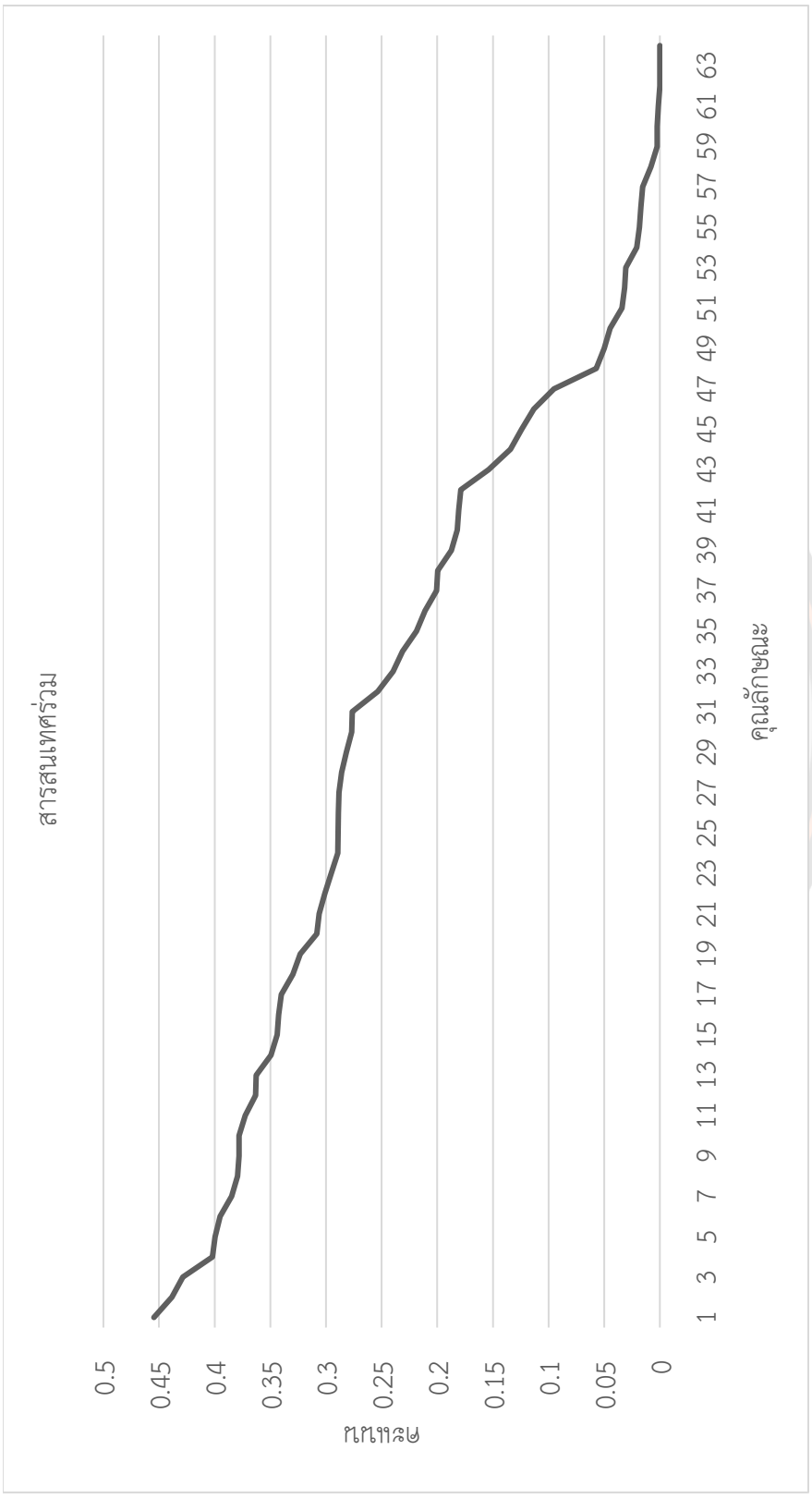
ภาพ **18** พบว่าค่าคะแนนมีพฤติกรรมที่ค่อยๆ ลดหลั่นลงไป ซึ่งสามารถระบุจุดแบ่งความสำคัญของคุณลักษณะได้ยาก สะท้อนถึงการที่ควรจะพิจารณาคุณลักษณะทั้งหมดในการค้นหาค่าไฮเปอร์พารามิเตอร์ อย่างไรก็ดีข้อสรุปนี้เป็นเพียงแค่แนวคิดแบบศึกษาสำนึก มิได้เป็นข้อบังคับแต่อย่างใด



ภาพ 16 ค่าคะแนนของคุณลักษณะด้วยวิธีไคสแควร์ของชุดข้อมูลลายมือ



ภาพ 17 ค่าคะแนนของคุณลักษณะด้วยวิธีการวิเคราะห์ความแปรปรวนของชุดข้อมูลลายมือ



ภาพ 18 ค่าคะแนนของคุณลักษณะด้วยวิธีสารสนเทศร่วมของชุดข้อมูลลายมือ

4.4 ผลการทดลองของขั้นตอนวิธีเชิงพันธุกรรม

เนื่องจากในวิทยานิพนธ์นี้ให้ความสำคัญกับการลดขนาดล็อกไฟล์สำหรับงานด้านความปลอดภัยทางเครือข่าย ในหัวข้อนี้จึงขอจำกัดการทดลองที่ชุดข้อมูลไอโอทีและเอ็นเอสแอลเคดีดีเท่านั้น โดยตารางผลลัพธ์จะแสดงให้สอดคล้องกับสองกรอบความคิดที่ได้อธิบายไว้ข้างต้น

ตาราง 32 และตาราง 33 แสดงตารางผลลัพธ์สำหรับชุดข้อมูลไอโอทีเรียงลำดับตามค่าความถูกต้องจากมากไปน้อย และเรียงลำดับตามจำนวนคุณลักษณะจากน้อยไปมาก เพื่อตอบโจทย์ของผู้ใช้ทั้งสองกรอบความคิด โดยพบว่าการเลือกคุณลักษณะจำนวน 35 ตัวแล้วนำไปสร้างคุณลักษณะพหุนามดีกรี 2 ตอบสนองความต้องการทั้งสองกรอบความคิด

ตาราง 34 และตาราง 35 แสดงตารางผลลัพธ์สำหรับชุดข้อมูลเอ็นเอสแอลเคดีดี ได้คำตอบไปในทิศทางเดียวกันทั้งสองกรอบความคิด คือการเลือกคุณลักษณะจำนวน 51 ตัว โดยไม่ต้องผ่านการสร้างคุณสมบัติพหุนาม

จากการทดลองทั้งสองจะเห็นได้ว่าขั้นตอนวิธีเชิงพันธุกรรมมีความยืดหยุ่นในการเลือกคุณลักษณะมากขึ้น ทำให้มีโอกาสได้ค่าความแม่นยำที่สูงขึ้น อย่างไรก็ดีจากการทดลองจะเห็นว่าจำนวนคุณลักษณะที่ถูกเลือกมีจำนวนประมาณครึ่งหนึ่งสอดคล้องกับการสุ่มเลือกหรือไม่เลือกคุณลักษณะที่คล้ายกับการโยนเหรียญ ซึ่งมีค่ามากกว่าจำนวนคุณลักษณะจากเฟรมเวิร์คการเลือกคุณสมบัติสองขั้นตอน ดังนั้นจึงขอเสนอแนวทางการปรับปรุงจากข้อจำกัดนี้ในบทถัดไป

ตาราง 32 ผลการทดลองของขั้นตอนวิธีเชิงพันธุกรรมของข้อมูลไอโอทีแบบเรียงตามค่าความถูกต้องจากมากไปน้อย

รุ่นที่	ดีกรี	จำนวนคุณลักษณะ	ค่าความถูกต้อง
54	2	44	0.953
63	2	52	0.953
82	2	45	0.953
9	2	41	0.953
12	2	37	0.953
13	2	43	0.953
17	2	42	0.953
22	2	45	0.953
23	2	43	0.953
30	2	35	0.953
45	2	43	0.953
46	2	43	0.953
47	2	45	0.953
48	2	50	0.953
52	2	43	0.953
61	2	46	0.953
14	2	41	0.953
16	2	42	0.953
20	2	45	0.953
34	2	40	0.953

ตาราง 33 ผลการทดลองของขั้นตอนวิธีเชิงพันธุกรรมของข้อมูลไอโอทีแบบเรียงตามจำนวนคุณลักษณะจากน้อยไปมาก

รุ่นที่	ดีกรี	จำนวนคุณลักษณะ	ค่าความถูกต้อง
30	2	35	0.953
21	2	35	0.952
100	2	35	0.952
33	2	36	0.952
32	2	36	0.952
96	2	36	0.952
74	2	36	0.952
12	2	37	0.953
5	2	37	0.952
3	2	37	0.952
15	2	37	0.952
31	2	37	0.952
35	2	37	0.952
73	2	37	0.952
2	2	37	0.952
97	2	37	0.952
51	2	39	0.953
29	2	39	0.952
38	2	39	0.952
1	2	39	0.952

ตาราง 34 ผลการทดลองของขั้นตอนวิธีเชิงพันธุกรรมของเอ็นเอสแอลเคดีดีแบบเรียงตามความถูกต้องแบบมากไปน้อย

รุ่นที่	ดีกรี	จำนวนคุณลักษณะ	ค่าความถูกต้อง
78	1	51	0.808
5	1	64	0.802
56	1	61	0.801
6	2	58	0.801
52	1	60	0.799
13	1	55	0.797
51	1	57	0.797
50	1	55	0.796
80	1	59	0.796
54	2	55	0.796
79	1	55	0.795
87	2	56	0.795
7	1	64	0.794
96	2	57	0.793
9	1	62	0.792
58	1	69	0.792
46	1	70	0.792
53	2	59	0.791
45	1	65	0.790
44	1	64	0.789

ตาราง 35 ผลการทดลองของขั้นตอนวิธีเชิงพันธุกรรมของข้อมูลเอ็นเอสแอลเคดีดีแบบเรียงตามจำนวนคุณลักษณะจากน้อยไปมาก

รุ่นที่	ดีกรี	จำนวนคุณลักษณะ	ค่าความถูกต้อง
78	1	51	0.808
71	1	52	0.777
8	1	54	0.759
49	1	54	0.785
75	1	54	0.783
13	1	55	0.797
20	1	55	0.782
31	1	55	0.765
50	1	55	0.796
54	2	55	0.796
61	1	55	0.776
79	1	55	0.795
23	1	56	0.781
43	1	56	0.772
55	1	56	0.788
74	1	56	0.777
77	1	56	0.774
87	2	56	0.795
2	1	57	0.774
10	1	57	0.784

4.5 ความสัมพันธ์ระหว่างจำนวนคุณลักษณะและขนาดของล็อกไฟล์

เนื่องด้วยกระบวนการเลือกคุณลักษณะถูกกำหนดด้วยจำนวนคุณลักษณะซึ่งในบางกรณีของชุดข้อมูลอาจจะไม่ได้มีความสัมพันธ์แบบเป็นเชิงเส้นกับขนาดของล็อกไฟล์ดังตาราง 36 ถึง ตาราง 39 ที่แสดงความสัมพันธ์ของจำนวนคุณลักษณะและขนาดของล็อกไฟล์ และแสดงค่าที่ได้จากการวิเคราะห์การถดถอยเชิงเส้น โดยค่าที่ได้จากตาราง 36 และตาราง 38 ถูกนำมาแสดงเป็นแผนภูมิเส้นใน

ภาพ **19** และภาพ 20ตามลำดับ

เมื่อพิจารณาจากแผนภูมิพบว่าลักษณะของแผนภูมิเส้นใน

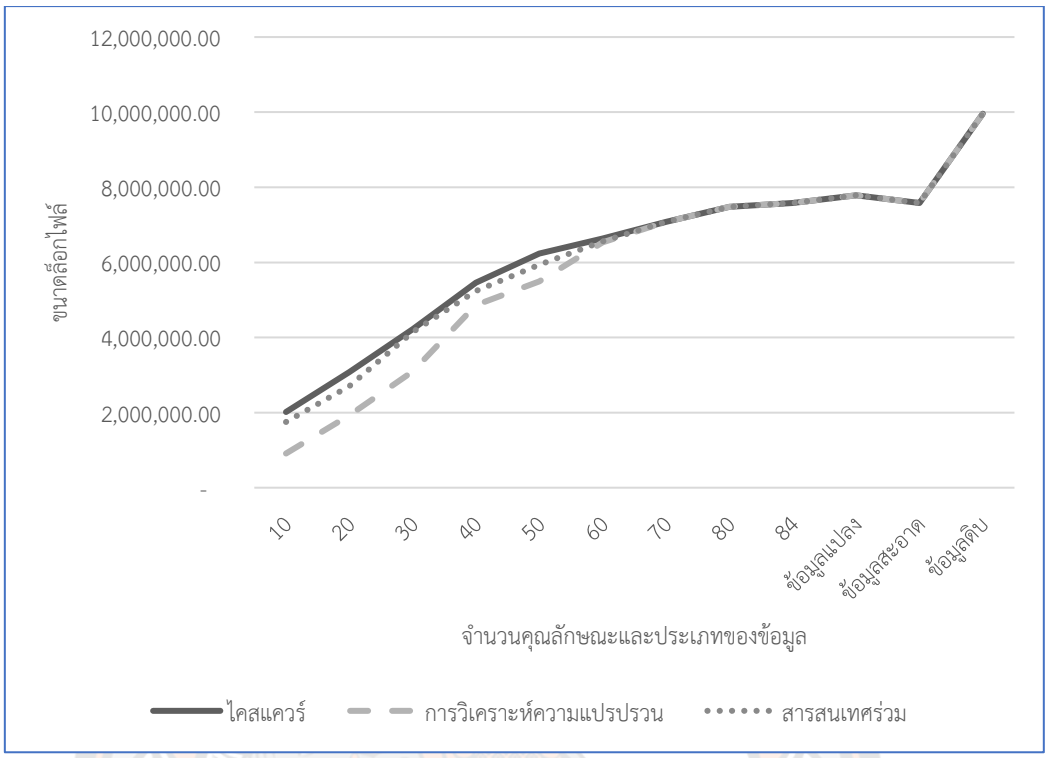
ภาพ **19** มีความใกล้เคียงกับความเป็นเชิงเส้น ในขณะที่ลักษณะของแผนภูมิเส้นในภาพ 20 มีความไม่ปกติเกิดขึ้นในกรณที่ประเภทข้อมูลเป็นแบบที่ถูกแปลงแล้ว ทั้งนี้เนื่องจากการเก็บล็อกไฟล์ที่ถูกแปลงแล้วโดยตรงทำให้เกิดกรณีที่คุณลักษณะที่เป็นหมวดหมู่มีจำนวนหมวดหมู่มาก เมื่อทำการแปลงข้อมูลแบบวันฮอทแล้วทำให้การจัดเก็บโดยตรงเกิดการใช้พื้นที่จัดเก็บเพิ่มมากขึ้นโดยไม่จำเป็น ยกตัวอย่างเช่น คุณลักษณะ flag ในชุดข้อมูลเอ็นเอสแอลเคดีดี มีจำนวนหมวดหมู่ 65 หมวดหมู่ ทุกหมวดหมู่จะถูกจัดเก็บเป็นคุณลักษณะใหม่ที่เพิ่มเข้ามา ดังนั้นในวิทยานิพนธ์นี้จึงเสนอแนะให้ทำการจัดเก็บล็อกไฟล์โดยที่คุณลักษณะที่เป็นหมวดหมู่ให้ทำการจัดเก็บในลักษณะเดิมโดยที่ไม่มีการแปลงข้อมูลแบบวันฮอทในขณะจัดเก็บ โดยแผนภูมิเส้นในภาพ 20 ส่วนที่แสดงความสัมพันธ์ระหว่างจำนวนคุณลักษณะและขนาดของล็อกไฟล์ได้ถูกปรับปรุงตามข้อเสนอแนะเพื่อแก้ปัญหาดังกล่าว

เมื่อได้การปรับปรุงวิธีการจัดเก็บล็อกไฟล์แล้ว งานวิจัยนี้ได้ทำการวิเคราะห์การถดถอยเชิงเส้นเพื่อศึกษาพฤติกรรมของความเปลี่ยนแปลงของจำนวนคุณลักษณะที่มีผลต่อขนาดของล็อกไฟล์ โดยพิจารณาข้อมูลเฉพาะส่วนที่เกี่ยวข้องกับจำนวนคุณลักษณะเท่านั้นดังแสดงผลลัพธ์ที่เป็นค่าต่างๆ จากการวิเคราะห์การถดถอยเชิงเส้นดังตาราง 37 และตาราง 39 ดังจะเห็นได้ว่าสำหรับชุดข้อมูลไอโอทีความสัมพันธ์ระหว่างจำนวนคุณลักษณะและขนาดของล็อกไฟล์ด้วยการทำการเลือกคุณลักษณะด้วยวิธีไคสแควร์ การวิเคราะห์ความแปรปรวน และสารสนเทศร่วม มีระดับของค่าอาร์สแควร์ที่สูงถึงประมาณ 95% ซึ่งค่าอาร์สแควร์ที่สูงสะท้อนถึงความถูกต้องของการวิเคราะห์การถดถอยเชิงเส้น ในขณะที่ความสัมพันธ์ที่ได้จากการวิเคราะห์ชุดข้อมูลเอ็นเอสแอลเคดีดี มีระดับของค่าอาร์สแควร์ที่ต่ำกว่าถึงระดับ 70% สะท้อนให้เห็นว่าการวิเคราะห์การถดถอยเชิงเส้นนั้นขาดความถูกต้อง และความสัมพันธ์ระหว่างจำนวนคุณลักษณะและขนาดของล็อกไฟล์ไม่เป็นเชิงเส้น ทั้งนี้เนื่องจากคุณลักษณะที่เป็นหมวดหมู่นั้นมีจำนวนมาก การจัดเก็บล็อกไฟล์โดยที่ไม่มีการแปลงข้อมูลประเภทหมวดหมู่ส่งผลให้ความสัมพันธ์ดังกล่าวมีความคลาดเคลื่อนเมื่อทำการวิเคราะห์การถดถอยเชิงเส้น

อย่างไรก็ดีด้วยผลลัพธ์ดังกล่าวแสดงให้เห็นว่าการลดจำนวนของคุณลักษณะด้วยวิธีการเลือกคุณลักษณะแบบสองขั้นตอนนั้นแปรผันตรงกับขนาดของล็อกไฟล์ และสามารถนำไปประยุกต์ใช้กับงานทางด้านการจัดเก็บล็อกไฟล์เพื่อการรักษาความปลอดภัยทางเครือข่ายได้

ตาราง 36 ความสัมพันธ์ของจำนวนคุณลักษณะและขนาดของล็อกไฟล์สำหรับชุดข้อมูลไอโอที

จำนวนคุณลักษณะ	ขนาดล็อกไฟล์ (ไบต์)		
	ไคสแควร์	การวิเคราะห์ความแปรปรวน	สารสนเทศร่วม
10	2,015,927	910,040	1,750,146
20	3,082,960	1,925,667	2,707,107
30	4,223,196	3,096,349	4,150,344
40	5,461,971	4,862,970	5,238,898
50	6,234,258	5,493,709	5,934,816
60	6,636,146	6,535,244	6,555,093
70	7,076,302	7,076,302	7,076,311
80	7,476,458	7,476,458	7,476,458
84	7,578,896	7,578,896	7,578,896
ข้อมูลแปลง	7,789,010	7,789,010	7,789,010
ข้อมูลสะอาด	7,578,896	7,578,896	7,578,896
ข้อมูลดิบ	9,959,178	9,959,178	9,959,178



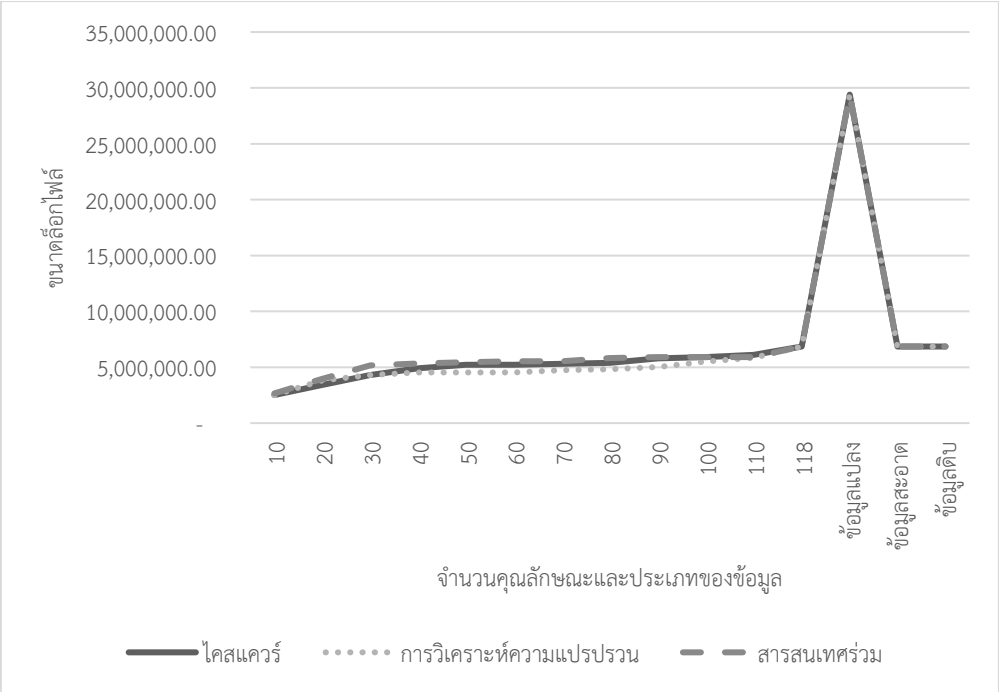
ภาพ 19 แผนภูมิแสดงความสัมพันธ์ของจำนวนคุณลักษณะและขนาดของล็อกไฟล์ของชุดข้อมูลไอโอที

ตาราง 37 ตารางแสดงค่าจากการวิเคราะห์การถดถอยเชิงเส้นของชุดข้อมูลไอโอที

	ค่าวิเคราะห์การถดถอยเชิงเส้น		
การให้คะแนน	ไคสแควร์	การวิเคราะห์ความแปรปรวน	สารสนเทศร่วม
จุดตัดแกน Y	1,881,150.998	426,967.232	1,522,284.293
ความชัน	73,999.448	92,596.689	78,305.204
ค่าอาร์สแควร์	0.946	0.961	0.955

ตาราง 38 ความสัมพันธ์ของจำนวนคุณลักษณะและขนาดของล็อกไฟล์สำหรับชุดข้อมูลเอ็นเอสแอลเคดีดี

จำนวนคุณลักษณะ	ขนาดล็อกไฟล์ (ไบต์)		
	ไคสแควร์	การวิเคราะห์ความแปรปรวน	สารสนเทศร่วม
10	2,514,961	2,507,715	2,657,493
20	3,417,311	3,815,881	3,959,846
30	4,291,594	4,313,781	5,157,673
40	4,912,534	4,522,288	5,348,853
50	5,216,543	4,522,288	5,444,125
60	5,216,543	4,522,288	5,539,766
70	5,312,184	4,726,450	5,539,812
80	5,407,669	4,829,035	5,826,270
90	5,802,796	5,020,003	5,921,800
100	5,921,816	5,497,533	5,921,689
110	6,112,776	5,840,535	5,921,796
118	6,860,762	6,860,762	6,860,762
ข้อมูลแปลง	29,401,419	29,401,419	29,401,419
ข้อมูลสะอาด	6,860,762	6,860,762	6,860,762
ข้อมูลดิบ	6,860,762	6,860,762	6,860,762



ภาพ 20 แผนภูมิแสดงความสัมพันธ์ของจำนวนคุณลักษณะและขนาดของล็อกไฟล์ของชุดข้อมูลไอโอที

ตาราง 39 ตารางแสดงค่าจากการวิเคราะห์การถดถอยเชิงเส้นของชุดข้อมูลเอ็นเอสแอลเคดีดี

	ค่าวิเคราะห์การถดถอยเชิงเส้น		
การให้คะแนน	ไคสแควร์	การวิเคราะห์ความแปรปรวน	สารสนเทศร่วม
จุดตัดแกน Y	3,059,458.284	2,972,214.201	3,669,567.614
ความชัน	31,200.501	27,393.302	25,790.583
ค่าอาร์สแควร์	0.875	0.843	0.732

4.6 ผลการประยุกต์ใช้ขนาดของล็อกไฟล์เพื่อการเลือกไฮเปอร์พารามิเตอร์และแบบจำลองที่เหมาะสมที่สุด

ตาราง 40 และตาราง 41 แสดงการคำนวณค่า $\boldsymbol{s_{max} - s_i, a_{max} - a_i}$ และค่าสัดส่วนการแลกเปลี่ยน $\frac{\boldsymbol{s_{max} - s_i}}{\boldsymbol{a_{max} - a_i}}$ โดยค่า $\boldsymbol{s_{max}}$ ถูกกำหนดให้เป็นค่าขนาดของล็อกไฟล์สะอาด

ในแถวของการค้นหาไฮเปอร์พารามิเตอร์แบบไม่ใช้การเลือกคุณลักษณะดังแสดงด้วยตัวเลขที่เป็นอักษรหนาและเอียง ขนาดของล็อกไฟล์ $\boldsymbol{S_i}$ จะเป็นค่าเดียวกันทั้งหมดสำหรับแต่ละวิธีการเลือกคุณลักษณะและจำนวนคุณลักษณะที่ถูกเลือก โดยไม่เปลี่ยนแปลงตามกระบวนการ M2, M3 หรือ M4 ค่าความแตกต่างของค่าความถูกต้องถูกคำนวณเทียบกับตาราง 28 และตาราง 29 ตามลำดับ จากผลลัพธ์ดังกล่าว สรุปได้ว่าสำหรับชุดข้อมูลไอโอที การเลือกคุณลักษณะที่มีค่าคะแนนสูงสุด 20 อันดับแรกด้วยวิธีการวิเคราะห์ความแปรปรวน เมื่อผ่านการสร้างคุณลักษณะพหุนามที่ดีกรี 3 แล้วนำไปค้นหาค่าไฮเปอร์พารามิเตอร์ จะให้ค่าที่เหมาะสุดที่เป็นการแลกเปลี่ยนระหว่างขนาดล็อกไฟล์ที่ลดลงและความถูกต้องที่ลดลง สำหรับชุดข้อมูลเอ็นเอสแอลเคดีดี การเลือกคุณลักษณะที่มีค่าคะแนนสูงสุด 20 อันดับแรกด้วยวิธีการวิเคราะห์ความแปรปรวน เมื่อผ่านการสร้างคุณลักษณะพหุนามที่ดีกรี 3 แล้วทำการเลือกคุณลักษณะพหุนามขั้นที่สองที่มีค่าคะแนนสูงสุด 20 อันดับแรก แล้วจึงนำไปค้นหาค่าไฮเปอร์พารามิเตอร์ จะให้ค่าที่เหมาะสุดของการแลกเปลี่ยน ผลลัพธ์ที่แสดงมาสะท้อนให้เห็นว่าการค้นหาค่าไฮเปอร์พารามิเตอร์ด้วยเฟรมเวิร์คที่นำเสนออาจจะไม่ได้ให้ค่าความถูกต้องที่สูงที่สุดเมื่อลดจำนวนคุณลักษณะและเพิ่มดีกรีความสัมพันธ์พหุนาม แต่สามารถให้การแลกเปลี่ยนที่คุ้มค่าที่สุดเมื่อมองทั้งกรอบของความถูกต้องและการลดลงของขนาดล็อกไฟล์

ตาราง 40 แสดงความแตกต่างของขนาดไฟล์และค่าสัดส่วนการเปลี่ยนแปลงของข้อมูลชุดไอโอที

วิธีการเลือกคุณลักษณะ	จำนวนคุณลักษณะ/ดีกรี	ขนาดล็อกไฟล์ (ไบต์)	ความแตกต่างของขนาดไฟล์	ความแตกต่างของค่าความถูกต้อง				ค่าสัดส่วนการแลกเปลี่ยน			
				M1	M2	M3	M4	M1	M2	M3	M4
ไม่ใช้	84/1	***7,578,896***		0.047				-			
ไคสแควร์	20/2	3,082,960	4,495,936		0.102	0.101	0.134		44,121,059.863	44,691,212.724	33,526,741.238
	20/3	3,082,960	4,495,936		0.101	0.098	0.121		44,646,832.175	45,970,715.746	37,248,848.384
	40/2	5,461,971	2,116,925		0.076	0.087	0.087		27,744,757.536	24,360,471.807	24,221,109.840
	40/3	5,461,971	2,116,925		0.078	0.091	0.091		27,140,064.103	23,365,618.102	23,288,503.850
การวิเคราะห์ความแปรปรวน	20/2	1,925,667	5,653,229		0.053	0.055	0.073		107,475,836.502	103,538,992.674	77,124,542.974
	20/3	1,925,667	5,653,229		0.052	0.058	0.079		**108,299,406.130**	97,469,465.517	71,832,642.948
	40/2	4,862,970	2,715,926		0.062	0.070	0.083		43,805,258.065	38,798,942.857	32,604,153.661
	40/3	4,862,970	2,715,926		0.064	0.074	0.098		42,370,140.406	36,504,381.720	27,628,952.187
สารสนเทศร่วม	20/2	2,707,107	4,871,789		0.081	0.071	0.086		59,923,603.936	68,713,526.093	56,714,656.577
	20/3	2,707,107	4,871,789		0.077	0.079	0.096		62,943,010.336	62,061,006.369	50,589,709.242
	40/2	5,238,898	2,339,998		0.064	0.070	0.076		36,734,662.480	33,238,607.955	30,952,354.497
	40/3	5,238,898	2,339,998		0.066	0.072	0.099		35,400,877.458	32,320,414.365	23,756,324.873

ตาราง 41 แสดงความแตกต่างของขนาดไฟล์และค่าสัดส่วนการเปลี่ยนแปลงของข้อมูลชุดเอ็นเอสแอลเคดีดี

วิธีการเลือกคุณลักษณะ	จำนวน คุณลักษณะ/ดีกรี	ขนาดล็อก ไฟล์ (ไบต์)	ความแตกต่าง ของขนาดไฟล์	ความแตกต่างของค่าความถูกต้อง				ค่าสัดส่วนการแลกเปลี่ยน			
				M1	M2	M3	M4	M1	M2	M3	M4
ไม่ใช้	118/1	***6,860,762***		0.225				-			
ไคสแควร์	20/2	3,417,311	3,443,451		0.215	0.213	0.213		16,001,166.357	16,151,271.107	16,189,238.364
	20/3	3,417,311	3,443,451		0.215	0.215	0.198		16,038,430.368	16,038,430.368	17,373,617.558
	40/2	4,912,534	1,948,228		0.223	0.218	0.229		8,748,217.333	8,932,728.106	8,526,161.926
	40/3	4,912,534	1,948,228		0.222	0.220	0.226		8,771,850.518	8,859,608.913	8,624,293.935
การวิเคราะห์ความแปรปรวน	20/2	3,815,881	3,044,881		0.269	0.263	0.283		11,315,053.883	11,590,715.645	10,778,339.823
	20/3	3,815,881	3,044,881		0.229	0.223	0.207		13,325,518.600	13,684,858.427	**14,738,049.371**
	40/2	4,522,288	2,338,474		0.242	0.238	0.253		9,659,124.329	9,842,062.290	9,228,389.897
	40/3	4,522,288	2,338,474		0.243	0.243	0.279		9,623,349.794	9,639,216.818	8,372,624.418
สารสนเทศร่วม	20/2	3,959,846	2,900,916		0.222	0.217	0.228		13,067,189.189	13,355,966.851	12,706,596.583
	20/3	3,959,846	2,900,916		0.221	0.219	0.196		13,126,316.742	13,228,071.135	14,838,445.013
	40/2	5,348,853	1,511,909		0.223	0.214	0.229		6,776,822.053	7,058,398.693	6,590,710.549
	40/3	5,348,853	1,511,909		0.223	0.213	0.225		6,785,947.038	7,094,833.412	6,731,562.778

บทที่ 5

บทสรุป

ในบทนี้จะขอกล่าวถึงข้อเสนอแนะพร้อมด้วยแนวทางในการปรับปรุงวิธีที่นำเสนอในวิทยานิพนธ์ รวมถึงบทสรุปของวิทยานิพนธ์เพื่อการประยุกต์ใช้และการพัฒนางานวิจัยในอนาคต

5.1 ข้อเสนอแนะ

5.1.1 การปรับปรุงการกำหนดจำนวนคุณลักษณะ

เฟรมเวิร์คการเลือกคุณลักษณะสองขั้นตอนที่นำเสนอมีข้อจำกัดในการกำหนดจำนวนคุณลักษณะที่ยังขาดการแนะนำผู้ใช้ให้เลือกจำนวนคุณลักษณะที่เหมาะสม แนวทางหนึ่งที่ขอเสนอคือการใช้ค่าระดับนัยสำคัญทางสถิติมาเป็นเกณฑ์ในการคัดกรองคุณลักษณะที่มีค่าความน่าจะเป็นที่ไม่สอดคล้องกับเกณฑ์ที่ได้จากค่าระดับนัยสำคัญทางสถิติออก ซึ่งเหตุการณ์ดังกล่าวไม่เกิดขึ้นกับชุดข้อมูลที่นำมาทดลองในวิทยานิพนธ์นี้ เนื่องจากค่าความน่าจะเป็นสำหรับคุณลักษณะที่มีค่ามากที่สุด 40 อันดับแรกมีค่าที่ต่ำมาก จึงไม่สามารถปฏิเสธสมมติฐานว่างได้ อย่างไรก็ดีสำหรับชุดข้อมูลอื่นที่มีจำนวนคุณลักษณะที่ไม่มีความสัมพันธ์กับค่าเป้าหมายเป็นจำนวนมาก สามารถประยุกต์ใช้วิธีการคัดกรองคุณลักษณะที่เสนอแนะนี้ได้ นอกจากนี้ยังจะสามารถใช้วิธีการหาจุดหักศอกบนค่าคะแนนเพื่อหาจุดวิกฤติในการกำหนดจำนวนของคุณลักษณะ ซึ่งวิธีนี้เป็นวิธีที่ได้รับความนิยมในการเลือกจำนวนกลุ่มของการจัดกลุ่มข้อมูลแบบที่ต้องระบุจำนวนข้อมูล

5.1.2 การปรับปรุงฟังก์ชันวัตถุประสงค์ของขั้นตอนวิธีเชิงพันธุกรรม

ขั้นตอนวิธีเชิงพันธุกรรมที่นำเสนอในวิทยานิพนธ์นี้พิจารณาเพียงแค่ค่าความถูกต้องในฟังก์ชันวัตถุประสงค์ โดยที่การลดลงของจำนวนคุณลักษณะถูกโน้มนำไปตามการสุ่มค่า 0 และ 1 แบบยูนิฟอร์ม จึงส่งผลให้จำนวนคุณลักษณะที่ถูกเลือกมีโอกาสที่จะอยู่ที่ประมาณครึ่งหนึ่งของจำนวนทั้งหมด เพื่อที่จะกำหนดแนวทางให้ขั้นตอนวิธีเชิงพันธุกรรมมีแนวโน้มที่จะลดจำนวนของคุณลักษณะลงจึงขอเสนอแนวทางการปรับปรุงฟังก์ชันวัตถุประสงค์ โดยการเพิ่มฟังก์ชันการลงโทษที่เป็นค่าติดลบที่สัมพันธ์กับจำนวนของคุณลักษณะ ซึ่งการปรับปรุงนี้จะขอนำไปเป็นงานวิจัยต่อยอดในอนาคต

5.2 บทสรุปวิทยานิพนธ์

วิทยานิพนธ์นี้นำเสนอเฟรมเวิร์คการเลือกคุณลักษณะสองขั้นตอนและ ขั้นตอนวิธีการคัดเลือกคุณลักษณะข้อมูลที่ใช้เทคนิคพหุนามร่วมกับขั้นตอนวิธีเชิงพันธุกรรม โดยในงานวิจัยนี้ได้รวม 3 เทคนิค ประกอบด้วย การคัดเลือกคุณลักษณะของชุดข้อมูล, การสร้างคุณลักษณะพหุนามและการปรับแต่งไฮเปอร์พารามิเตอร์ จากผลการทดลองที่ได้ตารางผลลัพธ์ของผลการทดลอง ได้นำตารางผลลัพธ์ มาพิจารณาความต้องการของผู้ใช้ ในสองมุมจากคำตอบที่ได้ พบว่า การทดลองจาก 4 ชุดข้อมูลสาธารณะที่เกี่ยวข้องกับประเภทการจำแนก มี 2 ชุดข้อมูลที่เป็นมาตรฐานและได้รับความนิยมในการทำงาน ทางด้านความปลอดภัยของเครือข่าย และอีก 2 ชุดข้อมูลถูกนำมาเพื่อแสดงความทั่วไปของเฟรมเวิร์ค และข้อจำกัดของเฟรมเวิร์คซึ่งจากการทดลองแบบต่างๆ พบว่า คำตอบที่ได้ผลลัพธ์ความแม่นยำกระจายตัวอยู่ทุก ๆ ขั้นตอนของเฟรมเวิร์คดังนั้นสรุปได้ว่า เฟรมเวิร์คมีความเป็นทั่วไป มีความใหม่ เป็นแนวคิดใหม่ และเปิดโอกาสให้ผู้ใช้ได้เลือกตัวเหมาะสมสำหรับการใช้งาน ทั้งในมุมของการจัดการจัดเก็บ แล้วก็ในมุมมองของความแม่นยำของของโมเดล แต่เฟรมเวิร์ค ยังมีข้อจำกัดเรื่องจำนวนของคุณลักษณะ ต้องถูกกำหนดจำนวนคุณลักษณะตั้งแต่แรก ทำให้ผู้ใช้งาน มีตัวเลือกน้อยลง ดังนั้น จึงได้นำเสนอวิธี ขั้นตอนวิธีการคัดเลือกคุณลักษณะข้อมูลที่ใช้เทคนิคพหุนามร่วมกับขั้นตอนวิธีเชิงพันธุกรรม โดยที่ วิธีนี้ ยังคงยึดหลักการเดิมแต่จะไม่จำกัดจำนวนของคุณลักษณะที่นำมาใช้งาน จากผลลัพธ์แสดงให้เห็นว่าขั้นตอนวิธีการคัดเลือกคุณลักษณะข้อมูลที่ใช้เทคนิคพหุนามร่วมกับขั้นตอนวิธีเชิงพันธุกรรม ได้ผลลัพธ์ที่ได้ความแม่นยำสูง และสามารถลดจำนวนคุณลักษณะพร้อมลดขนาดไฟล์และลดจำนวนคุณลักษณะ ขั้นตอนวิธีการคัดเลือกคุณลักษณะข้อมูลที่ใช้เทคนิคพหุนามร่วมกับขั้นตอนวิธีเชิงพันธุกรรม วิธีนี้ทำงาน 3 เทคนิคได้พร้อมกัน สามารถทำงาน ตั้งแต่การคัดเลือกคุณลักษณะของชุดข้อมูล การสร้างคุณลักษณะพหุนามและการปรับเปลี่ยนพารามิเตอร์ทำให้การทำงานมีประสิทธิภาพขึ้น

บรรณานุกรม

บรรณานุกรม

Aalaei, S., Shahraki, H., Rowhanimanesh, A., & Eslami, S. (2016). Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets. *Iranian journal of basic medical sciences*, *19*(5), 476.

Abdalla, R. R., & Jumaa, A. K. (2022). Log File Analysis Based on Machine Learning: A Survey: Survey. *UHD Journal of Science and Technology*, *6*(2), 77-84.

Abd-Alsabour, N. (2018). On the Role of Dimensionality Reduction. *J. Comput.*, *13*(5), 571-579.

Alduailij, M., Khan, Q. W., Tahir, M., Sardaraz, M., Alduailij, M., & Malik, F. (2022). Machine-Learning-Based DDoS Attack Detection Using Mutual Information and Random Forest Feature Importance Method. *Symmetry*, *14*(6), 1095.

Amini, F., & Hu, G. (2021). A two-layer feature selection method using genetic algorithm and elastic net. *Expert Systems with Applications*, *166*, 114072.

Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., & Lang, M. (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis*, *143*, 106839.

Brandao, A., & Georgieva, P. (2020, August). Log Files Analysis For Network Intrusion Detection. In *2020 IEEE 10th International Conference on Intelligent Systems (IS)* (pp. 328-333). IEEE.

David Akande, T., Kaur, B., Dadkhah, S., & Ghorbani, A. A. (2022, March). Threshold based technique to detect anomalies using log files. In *2022 7th International Conference on Machine Learning Technologies (ICMLT)* (pp. 191-198).

Eesa, A. S., Orman, Z., & Brifcani, A. M. A. (2015). A new feature selection model based on ID3 and bees algorithm for intrusion detection system. *Turkish Journal of Electrical Engineering and Computer Sciences*, *23*(2), 615-622.

Ertam, F., & Kaya, M. (2018, March). Classification of firewall log files with multiclass support vector machine. In *2018 6th International symposium on digital forensic and security (ISDFS)* (pp. 1-4). IEEE.

Gharaee, H., & Hosseinvand, H. (2016, September). A new feature selection IDS based on genetic algorithm and SVM. In *2016 8th International Symposium on Telecommunications (IST)* (pp. 139-144). IEEE.

Halimaa, A., & Sundarakantham, K. (2019, April). Machine learning based intrusion detection system. In *2019 3rd International conference on trends in electronics and informatics (ICOEI)* (pp. 916-920). IEEE.

Jain, D., & Singh, V. (2018). Feature selection and classification systems for chronic disease prediction: A review. *Egyptian Informatics Journal*, *19*(3), 179-189.

Jindal, P., & Kumar, D. (2017). A review on dimensionality reduction techniques. *Int. J. Comput. Appl*, *173*(2), 42-46.

Khotimah, B. K., Miswanto, M., & Suprajitno, H. (2020). Optimization of feature selection using genetic algorithm in naïve Bayes classification for incomplete data. *Int. J. Intell. Eng. Syst*, *13*(1), 334-343.

Meena Siwach, D. S. M. (2022). Anomaly detection for web log data analysis: A review. *Journal of Algebraic Statistics*, *13*(1), 129-148.

Nasiri, H., & Alavi, S. A. (2022). A novel framework based on deep learning and ANOVA feature selection method for diagnosis of COVID-19 cases from chest X-ray images. *Computational intelligence and neuroscience*, *2022*.

Ostertagová, E. (2012). Modelling using polynomial regression. *Procedia Engineering*, *48*, 500-506.

Oswald, C., Peichl, A., & Vyhlídal, T. (2021, June). Tensor-based Polynomial Features Generation for High-order Neural Networks. In *2021 23rd International Conference on Process Control (PC)* (pp. 175-179). IEEE.

Ritchey, R. P., & Perry, R. (2021, May). Machine Learning Toolkit for System Log File Reduction and Detection of Malicious Behavior. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (pp. 1-2). IEEE.

Ryciak, P., Wasielewska, K., & Janicki, A. (2022). Anomaly detection in log files using selected natural language processing methods. *Applied Sciences*, *12*(10), 5089.

Sciavicco, G., Zavatteri, M., & Villa, T. (2021). Mining CSTNUDs significant for a set of traces is polynomial. *Information and Computation*, *281*, 104773.

Sharma, N., & Saroha, K. (2015, May). Study of dimension reduction methodologies in data mining. In *International Conference on Computing, Communication & Automation* (pp. 133-137). IEEE.

Shekhawat, A. S., Di Troia, F., & Stamp, M. (2019). Feature analysis of encrypted malicious traffic. *Expert Systems with Applications*, *125*, 130-141.

Song, X. F., Zhang, Y., Gong, D. W., & Sun, X. Y. (2021). Feature selection using bare-bones particle swarm optimization with mutual information. *Pattern Recognition*, *112*, 107804.

Tama, B. A., Comuzzi, M., & Rhee, K. H. (2019). TSE-IDS: A two-stage classifier ensemble for intelligent anomaly-based intrusion detection system. *IEEE access*, *7*, 94497-94507.

Thaseen, I. S., Kumar, C. A., & Ahmad, A. (2019). Integrated intrusion detection model using chi-square feature selection and ensemble of classifiers. *Arabian Journal for Science and Engineering*, *44*, 3357-3368.

Velliangiri, S., & Alagumuthukrishnan, S. J. P. C. S. (2019). A review of dimensionality reduction techniques for efficient computation. *Procedia Computer Science*, *165*, 104-111.

Vinaroz, M., Charusaie, M. A., Harder, F., Adamczewski, K., & Park, M. J. (2022, June). Hermite polynomial features for private data generation. In *International Conference on Machine Learning* (pp. 22300-22324). PMLR

Viola, L., Ronchieri, E., & Cavallaro, C. (2022). Combining Log Files and Monitoring Data to Detect Anomaly Patterns in a Data Center. *Computers*, *11*(8), 117.

Wadekar, A., Gupta, T., Vijan, R., & Kazi, F. (2019, July). Hybrid CAE-VAE for unsupervised anomaly detection in log file systems. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-7). IEEE.

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A comprehensive review of dimensionality reduction techniques for feature

selection and feature extraction. *Journal of Applied Science and Technology Trends*, *1*(2), 56-70.

Zhao, M., Fu, C., Ji, L., Tang, K., & Zhou, M. (2011). Feature selection and parameter optimization for support vector machines: A new approach based on genetic algorithm with feature chromosomes. *Expert Systems with Applications*, *38*(5), 5197-5204.

Zhu, X., Wang, Z., Lin, P., Ma, Z., & Yu, Z. (2022, August). Data secrecy and security in computer network based on genetic algorithm. In *2022 International Conference on Artificial Intelligence in Everything (AIE)* (pp. 1-6). IEEE.

อภิธานศัพท์

อภิธานศัพท์

ลำดับ	คำศัพท์ (ไทย)	คำศัพท์ (อังกฤษ)
1.	การวิเคราะห์ความแปรปรวน	Analysis of variance (ANOVA)
2.	สารสนเทศร่วม	Mutual information
3.	พหุนาม	Polynomial
4.	คุณลักษณะ	Feature / Attribute
5.	ขั้นตอนวิธีเชิงพันธุกรรม	Genetic Algorithms
6.	ไคสแควร์	Chi-squared
7.	ชุดข้อมูลเอ็นเอสแอลเคดีดี	Nsl-kdd dataset
8.	ชุดข้อมูลไอโอที	IoT network intrusion dataset
9.	ชุดข้อมูลมะเร็ง	Breast cancer dataset
10.	ชุดข้อมูลลายมือ	Handwritten digit dataset
11.	ล็อกไฟล์	Log file
12.	การเลือกคุณลักษณะ	Feature Selection
13.	เฟรมเวิร์ค	Framework
14.	การนำเข้าชุดข้อมูล	Data Import
15.	การทำความสะอาดชุดข้อมูล	Data Cleansing
16.	การแปลงข้อมูล	Data Transformation
17.	แถว	Row
18.	คอลัมน์	Column
19.	ชุดข้อมูล	Dataset
20.	ข้อมูลอ้างอิง	Reference Data
21.	ข้อมูล	Data
22.	วิศวกรรมข้อมูล	Data Engineering
23.	การสำรวจข้อมูล	Data Exploration
24.	การนำเข้าชุดข้อมูล	Data Import

ลำดับ	คำศัพท์ (ไทย)	คำศัพท์ (อังกฤษ)
25.	การเรียนรู้ของเครื่อง, การเรียนรู้ด้วยตัวเองของคอมพิวเตอร์	Machine Learning
26.	แปลง, เปลี่ยนแปลง	Transform
27.	ชื่อพารามิเตอร์	Parameter name
28.	คำอธิบาย	Description
29.	การหาตัวแบบที่เหมาะสม	Model Fitting
30.	ทาเก็ต	Target
31.	ค่าความถูกต้อง	accuracy score
32.	ไฮเปอร์พารามิเตอร์	Hyperparameter
33.	การสร้างคุณลักษณะพหุนาม	Polynomial feature generation
34.	ไฮเปอร์พารามิเตอร์ลิสต์	Hyperparameter lists
35.	ข้อมูลฝึกสอน	Training set
36.	ข้อมูลทดสอบ	Test set
37.	โหนดใบ	Leaf Node
38.	แรนดอมฟอร์เรสต์	Random Forest
39.	การเข้ารหัสแบบวันฮอท	One-Hot encoding
40.	ทาร์เก็ต	Target
41.	ความสำคัญของคุณลักษณะ	Feature Importance
42.	การบุกรุกทางเครือข่าย	Intrusion Detection System; IDS
43.	การทำความสะอาดข้อมูลและแปลงข้อมูล	Cleansing and transformation
44.	การปรับไฮเปอร์พารามิเตอร์และการหาโมเดลที่เหมาะสม	Hyperparameter tuning and model fitting
45.	ไซคิทเลิร์น	Scikit-learn
46.	ประเภท	Class
47.	การวิเคราะห์องค์ประกอบหลัก, การแฮชคุณลักษณะ	PCA
48.	ตารางบันทึกคุณลักษณะพหุนาม	Polynomial feature lookup table

ลำดับ	คำศัพท์ (ไทย)	คำศัพท์ (อังกฤษ)
49.	อัลกอริธึม	algorithm
50.	การจัดลำดับความสำคัญของคุณลักษณะ	Recursive Feature Elimination (RFE)
51.	จำนวนบุคคล (โครโมโซม) ในแต่ละรุ่นของอัลกอริทึมทางพันธุกรรม	Population Size
52.	ความน่าจะเป็นที่ยีนในโครโมโซมจะกลายพันธุ์ในระหว่างกระบวนการวิวัฒนาการ	Mutation Rate
53.	ความน่าจะเป็นที่โครโมโซมพ่อแม่อยู่ระหว่างการผสมข้ามเพื่อผลิตโครโมโซมลูก	Crossover Rate
54.	จำนวนรุ่นสูงสุดที่อัลกอริธึมเชิงพันธุกรรมจะวนซ้ำ	Number of Generations
55.	วิธีการที่ใช้ในการเลือกโครโมโซมพ่อแม่เพื่อการสืบพันธุ์	Selection Method
56.	ฟังก์ชั่นวัตถุประสงค์ที่ใช้ในการประเมินสมรรถภาพ	Fitness Function
57.	รุ่น	Generation
58.	ฟังก์ชันการลงโทษ	Penalty function
59.	วิธีการเลือกคุณลักษณะแบบสองขั้นตอน	Two state
60.	แมชชีนเลิร์นนิ่งอัลกอริทึม	Machine Learning Algorithm
61.	ฝึกสอนโมเดล	Train Model
62.	ไฮเปอร์พารามิเตอร์ออฟติมัย	Hyperparameter Optimization
63.	ปัญหาการหาค่าที่เหมาะสมที่สุด	Optimization Problem
64.	เซต	Set
65.	การปรับค่าชุดไฮเปอร์พารามิเตอ	Traditional Hyperparameter Tuning

ลำดับ	คำศัพท์ (ไทย)	คำศัพท์ (อังกฤษ)
	ร์ที่เหมาะสม โดยตนเอง	
66.	การปรับค่าชุดไฮเปอร์พารามิเตอร์ที่เหมาะสมโดยอัตโนมัติ	Automated Hyperparameter Tuning
67.	อัลกอริธึมของดิซิชั่นทรี	Algorithm of Decision Tree
68.	วิธีการห่อหุ้ม	Wrapper Method
69.	วิธีการกรอง	Filter Method
70.	ค่าความน่าจะเป็น	Probability value: p-value
71.	นัยสำคัญทางสถิติ	Statistical significance
72.	สมมติฐานว่าง	Null hypothesis: H0
73.	โอเวอร์ฟิต	overfitting
74.	วิธีการฝังตัว	Embedded Method
75.	วิธีการผสาน	Hybrid Method
76.	วิธีการแบบรวม	Ensemble Method
77.	การเข้ารหัสโครโมโซม	Chromosome encoding
78.	กระบวนการทางพันธุกรรม	Genetic Operation
79.	การสลับสายพันธุ์	Crossover
80.	การกลายพันธุ์	Mutation
81.	โครโมโซมลูก	Offspring
82.	การสลับสายพันธ์แบบจุดเดียว	One-point crossover
83.	การคำนวณค่าความเหมาะสม	Fitness Computation
84.	การคัดเลือกสายพันธุ์	Selection
85.	โครโมโซม	Chromosome
86.	เข้ารหัส	Encoding
87.	ตัวเลขฐานสอง	Binary bit string
88.	ประชากร	Population
89.	การคัดเลือกแบบการจัดอันดับ	Linear ranking
90.	การคัดเลือกแบบการแข่งขัน	Tournament
91.	การคัดเลือกแบบวงล้อรูเล็ต	Roulette wheel
92.	ขั้นตอนวิธีทางพันธุกรรมอย่าง	Simple Genetic algorithm/SGA

ลำดับ	คำศัพท์ (ไทย)	คำศัพท์ (อังกฤษ)
	ง่าย	
93.	การตรวจสอบเงื่อนไขการสิ้นสุดการทำงาน	Termination condition

ภาคผนวก

ตาราง 42 แสดงรายละเอียดของชุดข้อมูลไอโอที

ลำดับคุณลักษณะ	ชื่อคุณลักษณะ	คำอธิบาย	ประเภทของข้อมูล
1	Flow_ID	รหัสโฟลว	ค่าตัวเลข
2	Src_IP	ที่อยู่ไอพีต้นทาง	ค่าสายอักขระ
3	Src_Port	หมายเลขพอร์ตต้นทาง	ค่าตัวเลข
4	Dst_IP	ที่อยู่ไอพีปลายทาง	ค่าสายอักขระ
5	Dst_Port	หมายเลขพอร์ตปลายทาง	ค่าตัวเลข
6	Protocol	โปรโตคอล	ค่าหมวดหมู่
7	Timestamp	เวลา	ค่าสายอักขระ
8	Flow_Duration	ระยะเวลาของโฟลว	ค่าตัวเลข
9	Tot_Fwd_Pkts	จำนวนแพ็กเก็ตทั้งหมดในทิศทางส่งไป	ค่าตัวเลข
10	Tot_Bwd_Pkts	จำนวนแพ็กเก็ตทั้งหมดในทิศทางย้อนกลับ	ค่าตัวเลข
11	TotLen_Fwd_Pkts	ขนาดรวมของแพ็กเก็ตในทิศทางส่งไป	ค่าตัวเลข
12	TotLen_Bwd_Pkts	ขนาดรวมของแพ็กเก็ตในทิศทางย้อนกลับ	ค่าตัวเลข
13	Fwd_Pkt_Len_Max	ขนาดสูงสุดของแพ็กเก็ตในทิศทางส่งไป	ค่าตัวเลข
14	Fwd_Pkt_Len_Min	ขนาดต่ำสุดของแพ็กเก็ตในทิศทางส่งไป	ค่าตัวเลข
15	Fwd_Pkt_Len_Mean	ขนาดเฉลี่ยของแพ็กเก็ตในทิศทาง	ค่าตัวเลข

ลำดับคุณลักษณะ	ชื่อคุณลักษณะ	คำอธิบาย	ประเภทของข้อมูล
		ส่งไป	
16	Fwd_Pkt_Len_Std	ค่าเบี่ยงเบนมาตรฐานของขนาดแพ็กเก็ตในทิศทางส่งไป	ค่าตัวเลข
17	Bwd_Pkt_Len_Max	ขนาดสูงสุดของแพ็กเก็ตในทิศทางย้อนกลับ	ค่าตัวเลข
18	Bwd_Pkt_Len_Min	ขนาดต่ำสุดของแพ็กเก็ตในทิศทางย้อนกลับ	ค่าตัวเลข
19	Bwd_Pkt_Len_Mean	ขนาดเฉลี่ยของแพ็กเก็ตในทิศทางย้อนกลับ	ค่าตัวเลข
20	Bwd_Pkt_Len_Std	ค่าเบี่ยงเบนมาตรฐานของขนาดแพ็กเก็ตในทิศทางย้อนกลับ	ค่าตัวเลข
21	Flow_Byts/s	อัตราบิตของโฟลวที่เป็นจำนวนแพ็กเก็ตที่ถ่ายโอนต่อวินาที	ค่าตัวเลข
22	Flow_Pkts/s	อัตราแพ็กเก็ตของโฟลวที่เป็นจำนวนแพ็กเก็ตที่ถ่ายโอนต่อวินาที	ค่าตัวเลข
23	Flow_IAT_Mean	เวลาเฉลี่ยระหว่างโฟลวสองรายการ	ค่าตัวเลข
24	Flow_IAT_Std	ค่าเบี่ยงเบนมาตรฐานของเวลาระหว่างโฟลวสองรายการ	ค่าตัวเลข
25	Flow_IAT_Max	เวลาสูงสุดระหว่างโฟลวสองรายการ	ค่าตัวเลข
26	Flow_IAT_Min	เวลาต่ำสุดระหว่างโฟลวสองรายการ	ค่าตัวเลข
27	Fwd_IAT_Tot	เวลาทั้งหมดระหว่างการส่งแพ็กเก็ตสองรายการในทิศทางส่งไป	ค่าตัวเลข
28	Fwd_IAT_Mean	เวลาเฉลี่ยระหว่างการส่งแพ็กเก็ต	ค่าตัวเลข

ลำดับคุณลักษณะ	ชื่อคุณลักษณะ	คำอธิบาย	ประเภทของข้อมูล
		สองรายการในทิศทางส่งไป	
29	Fwd_IAT_Std	ค่าเบี่ยงเบนมาตรฐานของเวลาระหว่างการส่งแพ็กเก็ตสองรายการในทิศทางส่งไป	ค่าตัวเลข
30	Fwd_IAT_Max	เวลาสูงสุดระหว่างการส่งแพ็กเก็ตสองรายการในทิศทางส่งไป	ค่าตัวเลข
31	Fwd_IAT_Min	เวลาต่ำสุดระหว่างการส่งแพ็กเก็ตสองรายการในทิศทางส่งไป	ค่าตัวเลข
32	Bwd_IAT_Tot	เวลาทั้งหมดระหว่างการส่งแพ็กเก็ตสองรายการในทิศทางย้อนกลับ	ค่าตัวเลข
33	Bwd_IAT_Mean	เวลาเฉลี่ยระหว่างการส่งแพ็กเก็ตสองรายการในทิศทางย้อนกลับ	ค่าตัวเลข
34	Bwd_IAT_Std	ค่าเบี่ยงเบนมาตรฐานของเวลาระหว่างการส่งแพ็กเก็ตสองรายการในทิศทางย้อนกลับ	ค่าตัวเลข
35	Bwd_IAT_Max	เวลาสูงสุดระหว่างการส่งแพ็กเก็ตสองรายการในทิศทางย้อนกลับ	ค่าตัวเลข
36	Bwd_IAT_Min	เวลาต่ำสุดระหว่างการส่งแพ็กเก็ตสองรายการในทิศทางย้อนกลับ	ค่าตัวเลข
37	Fwd_PSH_Flags	จำนวนครั้งที่ตรงกับแฟลก PSH ในแพ็กเก็ตที่เดินทางในทิศทางส่งไป (0 สำหรับยูดีพี)	ค่าตัวเลข
38	Bwd_PSH_Flags	จำนวนครั้งที่ตรงกับแฟลก PSH ในแพ็กเก็ตที่เดินทางในทิศทางย้อนกลับ (0 สำหรับยูดีพี)	ค่าตัวเลข

ลำดับคุณลักษณะ	ชื่อคุณลักษณะ	คำอธิบาย	ประเภทของข้อมูล
39	Fwd_URG_Flags	จำนวนครั้งที่ตรงกับแฟลก URG ในแพ็กเก็ตที่เดินทางในทิศทางส่งไป (0 สำหรับยูดีพี)	ค่าตัวเลข
40	Bwd_URG_Flags	จำนวนครั้งที่ตรงกับแฟลก URG ในแพ็กเก็ตที่เดินทางในทิศทางย้อนกลับ (0 สำหรับยูดีพี)	ค่าตัวเลข
41	Fwd_Header_Len	จำนวนไบต์รวมที่ใช้สำหรับข้อมูลส่วนหัวในทิศทางส่งไป	ค่าตัวเลข
42	Bwd_Header_Len	จำนวนไบต์รวมที่ใช้สำหรับข้อมูลส่วนหัวในทิศทางย้อนกลับ	ค่าตัวเลข
43	Fwd_Pkts/s	จำนวนแพ็กเก็ตทางส่งไปต่อวินาที	ค่าตัวเลข
44	Bwd_Pkts/s	จำนวนแพ็กเก็ตทางย้อนกลับต่อวินาที	ค่าตัวเลข
45	Pkt_Len_Min	ความยาวขั้นต่ำของโฟลว	ค่าตัวเลข
46	Pkt_Len_Max	ความยาวสูงสุดของโฟลว	ค่าตัวเลข
47	Pkt_Len_Mean	ความยาวเฉลี่ยของโฟลว	ค่าตัวเลข
48	Pkt_Len_Std	ค่าเบี่ยงเบนมาตรฐานของความยาวของโฟลว	ค่าตัวเลข
49	Pkt_Len_Var	เวลาระยะห่างขั้นต่ำระหว่างแพ็กเก็ต	ค่าตัวเลข
50	FIN_Flag_Cnt	จำนวนแพ็กเก็ตที่มีแฟลก FIN	ค่าตัวเลข
51	SYN_Flag_Cnt	จำนวนแพ็กเก็ตที่มีแฟลก SYN	ค่าตัวเลข
52	RST_Flag_Cnt	จำนวนแพ็กเก็ตที่มีแฟลก RST	ค่าตัวเลข

ลำดับคุณลักษณะ	ชื่อคุณลักษณะ	คำอธิบาย	ประเภทของข้อมูล
53	PSH_Flag_Cnt	จำนวนแพ็กเก็ตที่มีแฟลก PUSH	ค่าตัวเลข
54	ACK_Flag_Cnt	จำนวนแพ็กเก็ตที่มีแฟลก ACK	ค่าตัวเลข
55	URG_Flag_Cnt	จำนวนแพ็กเก็ตที่มีแฟลก URG	ค่าตัวเลข
56	CWE_Flag_Count	จำนวนแพ็กเก็ตที่มีแฟลก CWE	ค่าตัวเลข
57	ECE_Flag_Cnt	จำนวนแพ็กเก็ตที่มีแฟลก ECE	ค่าตัวเลข
58	Down/Up_Ratio	อัตราส่วนการดาวน์โหลดและอัปโหลด	ค่าตัวเลข
59	Pkt_Size_Avg	ขนาดเฉลี่ยของแพ็กเก็ต	ค่าตัวเลข
60	Fwd_Seg_Size_Avg	ขนาดเฉลี่ยที่สังเกตเห็นในทิศทางส่งไป	ค่าตัวเลข
61	Bwd_Seg_Size_Avg	ขนาดเฉลี่ยที่สังเกตเห็นในทิศทางย้อนกลับ	ค่าตัวเลข
62	Fwd_Byts/b_Avg	อัตราส่วนจำนวนไบต์กลุ่มในทิศทางส่งไป	ค่าตัวเลข
63	Fwd_Pkts/b_Avg	อัตราส่วนจำนวนแพ็กเก็ตกลุ่มในทิศทางส่งไป	ค่าตัวเลข
64	Fwd_Blk_Rate_Avg	อัตราส่วนจำนวนกลุ่มในทิศทางส่งไป	ค่าตัวเลข
65	Bwd_Byts/b_Avg	อัตราส่วนจำนวนไบต์กลุ่มในทิศทางย้อนกลับ	ค่าตัวเลข
66	Bwd_Pkts/b_Avg	อัตราส่วนจำนวนแพ็กเก็ตกลุ่ม	ค่าตัวเลข

ลำดับคุณลักษณะ	ชื่อคุณลักษณะ	คำอธิบาย	ประเภทของข้อมูล
		ในทิศทางย้อนกลับ	
67	Bwd_Blk_Rate_Avg	อัตราส่วนจำนวนกลุ่มในทิศทางย้อนกลับ	ค่าตัวเลข
68	Subflow_Fwd_Pkts	จำนวนเฉลี่ยของแพ็กเก็ตในการโฟลวย่อยในทิศทางส่งไป	ค่าตัวเลข
69	Subflow_Fwd_Byts	จำนวนเฉลี่ยของไบต์ในการโฟลวย่อยในทิศทางส่งไป	ค่าตัวเลข
70	Subflow_Bwd_Pkts	จำนวนเฉลี่ยของแพ็กเก็ตในการโฟลวย่อยในทิศทางย้อนกลับ	ค่าตัวเลข
71	Subflow_Bwd_Byts	จำนวนเฉลี่ยของไบต์ในการโฟลวย่อยในทิศทางย้อนกลับ	ค่าตัวเลข
72	Init_Fwd_Win_Byts	จำนวนไบต์ที่ส่งในหน้าต่างเริ่มต้นในทิศทางส่งไป	ค่าตัวเลข
73	Init_Bwd_Win_Byts	จำนวนไบต์รวมที่ส่งในหน้าต่างเริ่มต้นในทิศทางย้อนกลับ	ค่าตัวเลข
74	Fwd_Act_Data_Pkts	จำนวนแพ็กเก็ตรวมที่มีข้อมูลเพย์โหลดของข้อมูลทีซีพีอย่างน้อย 1 ไบต์ในทิศทางส่งไป	ค่าตัวเลข
75	Fwd_Seg_Size_Min	ขนาดเซกเมนต์ขั้นต่ำที่สังเกตได้ในทิศทางส่งไป	ค่าตัวเลข
76	Active_Mean	เวลาเฉลี่ยที่โฟลวก่อนแอคทีฟที่	ค่าตัวเลข

ลำดับคุณลักษณะ	ชื่อคุณลักษณะ	คำอธิบาย	ประเภทของข้อมูล
		จะเป็นไม่ได้ใช้งาน	
77	Active_Std	ค่าเบี่ยงเบนมาตรฐานของเวลาที่โฟลวก่อนแอคทีฟที่จะเป็นไม่ได้ใช้งาน	ค่าตัวเลข
78	Active_Max	เวลาสูงสุดที่โฟลวแอคทีฟก่อนที่จะไม่ได้ใช้งาน	ค่าตัวเลข
79	Active_Min	เวลาต่ำสุดที่โฟลวแอคทีฟก่อนที่จะเป็นไม่ได้ใช้งาน	ค่าตัวเลข
80	Idle_Mean	เวลาเฉลี่ยที่โฟลวว่างก่อนแอคทีฟที่จะเป็นใช้งานอีกครั้ง	ค่าตัวเลข
81	Idle_Std	ค่าเบี่ยงเบนมาตรฐานของเวลาที่โฟลวว่างก่อนแอคทีฟที่จะเป็นใช้งานอีกครั้ง	ค่าตัวเลข
82	Idle_Max	เวลาสูงสุดที่โฟลวว่างก่อนแอคทีฟที่จะเป็นใช้งานอีกครั้ง	ค่าตัวเลข
83	Idle_Min	เวลาต่ำสุดที่โฟลวว่างก่อนแอคทีฟที่จะเป็นใช้งานอีกครั้ง	ค่าตัวเลข

ตาราง 43 แสดงรายละเอียดของชุดข้อมูลเอ็นเอสแอลเคดีดี

ลำดับคุณลักษณะ	ชื่อคุณลักษณะ	คำอธิบาย	ประเภทของข้อมูล
1	duration	ระยะเวลาของการเชื่อมต่อ (วินาที)	ค่าตัวเลข

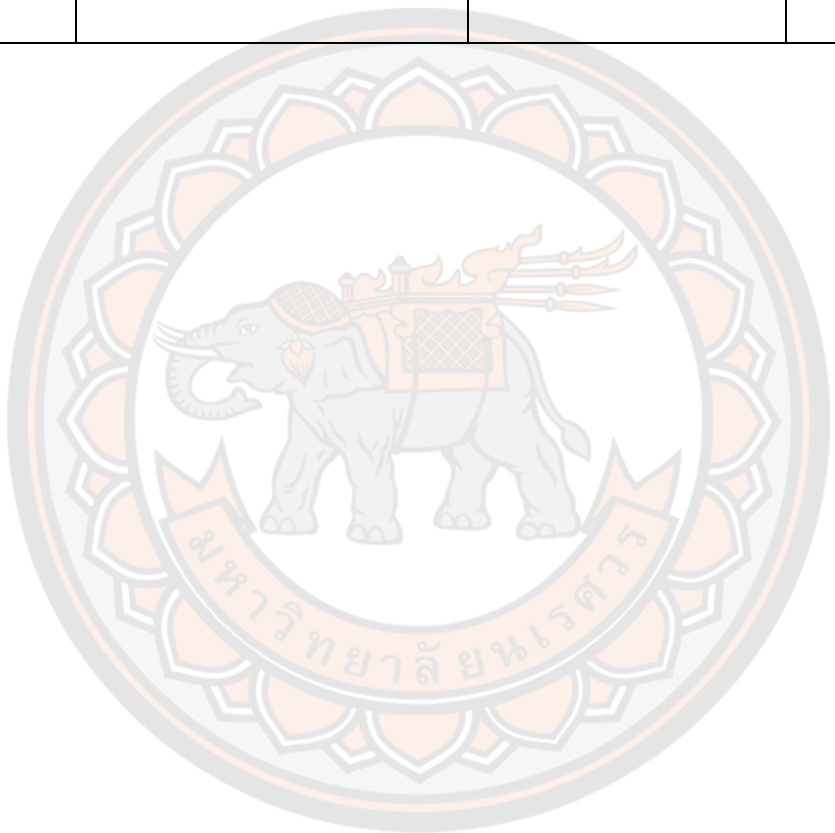
ลำดับคุณลักษณะ	ชื่อคุณลักษณะ	คำอธิบาย	ประเภทของข้อมูล
2	protocol_type	ประเภทของโปรโตคอล	ค่าหมวดหมู่
3	service	ประเภทของบริการทางเครือข่าย	ค่าหมวดหมู่
4	flag	ค่าความปกติหรือประเภทของความผิดพลาดของการเชื่อมต่อ	ค่าหมวดหมู่
5	src_bytes	จำนวนไบต์ที่ส่งจากต้นทางสู่ปลายทาง	ค่าตัวเลข
6	dst_bytes	จำนวนไบต์ที่ส่งจากปลายทางสู่ต้นทาง	ค่าตัวเลข
7	land	การเชื่อมต่อมาจากหรือไปยังโฮสต์และพอร์ตเดียวกันหรือไม่	ค่าไบนารี
8	wrong_fragment	จำนวนของส่วนย่อยที่ผิดพลาดในการเชื่อมต่อ	ค่าตัวเลข
9	urgent	จำนวนแพ็กเก็ตที่เร่งด่วนในการเชื่อมต่อ	ค่าตัวเลข
10	hot	จำนวนตัวบ่งชี้ 'hot'	ค่าตัวเลข
11	num_failed_logins	จำนวนความพยายามเข้าสู่ระบบที่ไม่สำเร็จ	ค่าตัวเลข
12	logged_in	ผู้ใช้เข้าสู่ระบบสำเร็จ	ค่าไบนารี

ลำดับคุณลักษณะ	ชื่อคุณลักษณะ	คำอธิบาย	ประเภทของข้อมูล
		หรือไม่	
13	num_compromised	จำนวนสภาวะที่ถูกโจมตี	ค่าตัวเลข
14	root_shell	ได้รับสิทธิ์รูทเชลหรือไม่	ค่าไบนารี
15	su_attempted	มีความพยายามใช้คำสั่ง 'su root' หรือไม่	ค่าไบนารี
16	num_root	จำนวนการเข้าถึงรูท	ค่าตัวเลข
17	num_file_creations	จำนวนการสร้างไฟล์	ค่าตัวเลข
18	num_shells	จำนวนการเรียกเชลพร้อมท์	ค่าตัวเลข
19	num_access_files	จำนวนการดำเนินการบนไฟล์ควบคุมการเข้าถึง	ค่าตัวเลข
20	num_outbound_cmds	จำนวนคำสั่งที่ส่งออกในเซสชันเอฟทีพี	ค่าตัวเลข
21	is_host_login	โฮสต์เข้าสู่ระบบสำเร็จหรือไม่	ค่าไบนารี
22	is_guest_login	ผู้เยี่ยมชมเข้าสู่ระบบสำเร็จหรือไม่	ค่าไบนารี
23	count	จำนวนการเชื่อมต่อไปยังโฮสต์เดียวกัน	ค่าตัวเลข
24	srv_count	จำนวนการเชื่อมต่อไปยังบริการเดียวกัน	ค่าตัวเลข

ลำดับคุณลักษณะ	ชื่อคุณลักษณะ	คำอธิบาย	ประเภทของข้อมูล
25	serror_rate	เปอร์เซ็นต์ของการเชื่อมต่อที่มีข้อผิดพลาด 'SYN'	ค่าตัวเลข
26	srv_serror_rate	เปอร์เซ็นต์ของการเชื่อมต่อที่มีข้อผิดพลาด 'SYN' (บริการ)	ค่าตัวเลข
27	rerror_rate	เปอร์เซ็นต์ของการเชื่อมต่อที่มีข้อผิดพลาด 'REJ'	ค่าตัวเลข
28	srv_rerror_rate	เปอร์เซ็นต์ของการเชื่อมต่อที่มีข้อผิดพลาด 'REJ' (บริการ)	ค่าตัวเลข
29	same_srv_rate	เปอร์เซ็นต์ของการเชื่อมต่อไปยังบริการเดียวกัน	ค่าตัวเลข
30	diff_srv_rate	เปอร์เซ็นต์ของการเชื่อมต่อไปยังบริการที่แตกต่างกัน	ค่าตัวเลข
31	srv_diff_host_rate	เปอร์เซ็นต์ของการเชื่อมต่อไปยังโฮสต์ที่แตกต่างกัน (บริการ)	ค่าตัวเลข
32	dst_host_count	จำนวนการเชื่อมต่อไปยังโฮสต์เดียวกัน (ปลายทาง)	ค่าตัวเลข

ลำดับคุณลักษณะ	ชื่อคุณลักษณะ	คำอธิบาย	ประเภทของข้อมูล
33	dst_host_srv_count	จำนวนการเชื่อมต่อไปยังบริการเดียวกัน (ปลายทาง)	ค่าตัวเลข
34	dst_host_same_srv_rate	เปอร์เซ็นต์ของการเชื่อมต่อไปยังบริการเดียวกัน (ปลายทาง)	ค่าตัวเลข
35	dst_host_diff_srv_rate	เปอร์เซ็นต์ของการเชื่อมต่อไปยังบริการที่แตกต่างกัน (ปลายทาง)	ค่าตัวเลข
36	dst_host_same_src_port_rate	เปอร์เซ็นต์ของการเชื่อมต่อจากพอร์ตต้นทางเดียวกัน (ปลายทาง)	ค่าตัวเลข
37	dst_host_srv_diff_host_rate	เปอร์เซ็นต์ของการเชื่อมต่อไปยังโฮสต์ที่แตกต่างกัน (บริการปลายทาง)	ค่าตัวเลข
38	dst_host_serror_rate	เปอร์เซ็นต์ของการเชื่อมต่อที่มีข้อผิดพลาด 'SYN' (ปลายทาง)	ค่าตัวเลข
39	dst_host_srv_serror_rate	เปอร์เซ็นต์ของการเชื่อมต่อที่มีข้อผิดพลาด 'SYN' (บริการปลายทาง)	ค่าตัวเลข
40	dst_host_rerror_rate	เปอร์เซ็นต์ของการเชื่อมต่อที่มีข้อผิดพลาด	ค่าตัวเลข

ลำดับ คุณลักษณะ	ชื่อคุณลักษณะ	คำอธิบาย	ประเภทของข้อมูล
		'REJ' (ปลายทาง)	
41	dst_host_srv_rerror_rate	เปอร์เซ็นต์ของการ เชื่อมต่อที่มีข้อผิดพลาด 'REJ' (บริการปลายทาง)	ค่าตัวเลข



ประวัติผู้วิจัย

ชื่อ-นามสกุล	กิตติภพ มหาวัน
วัน เดือน ปี เกิด	19 ตุลาคม 2516
ที่อยู่ปัจจุบัน	มหาวิทยาลัยราชภัฏพระนครศรีอยุธยา จังหวัดพระนครศรีอยุธยา
ที่ทำงานปัจจุบัน	อาจารย์ประจำ
ตำแหน่งหน้าที่ปัจจุบัน	พ.ศ. 2557 มหาวิทยาลัยราชภัฏพระนครศรีอยุธยา พ.ศ. 2556 มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา น่าน พ.ศ. 2542 มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา พิษณุโลก
ประสบการณ์การทำงาน	พ.ศ. 2549 วทม.(เทคโนโลยีสารสนเทศ) มหาวิทยาลัยนเรศวร พ.ศ. 2540 วทบ.(วิทยาการคอมพิวเตอร์) สถาบันราชภัฏอุตรดิตถ์
ประวัติการศึกษา	Kittiphop Mahawan,Winai Wongthai,and Surapong Wiriya. (2019).The Comparison of Electricity Usage Forecasting with Machinie Learning Techniques and Classical Statiscal Model.Proceeding of 4th National Conference and 2nd International Conference,Chaopraya University (pp.580-588). Nakhon Sawan: Chaopraya University