ระบบตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่อง สำหรับเสริมสร้างความแข็งแกร่งและมีประสิทธิภาพด้านการรักษาความมั่นคงปลอดภัยไซเบอร์ในองค์กร

ณัฐดนัย กองขุนทด[1*] ประสงค์ ปราณีตพลกรัง[2] พายัพ ศิรินาม[3]

[1,2,3]หลักสูตรวิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีป้องกันประเทศ สำนักบัณฑิตศึกษา*
โรงเรียนนายเรืออากาศนวมินทกษัตริยาธิราช, กรุงเทพมหานคร, ประเทศไทย

*ผู้ประพันธ์บรรณกิจ อีเมล : natdanai_k@rtaf.mi.th

รับต้นฉบับ : 13 เมษายน 2567; รับบทความฉบับแก้ไข : 12 พฤษภาคม 2567; ตอบรับบทความ : 28 พฤษภาคม 2567
เผยแพร่ออนไลน์ : 26 ธันวาคม 2567

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อสร้างระบบตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่อง สำหรับเสริมสร้างความแข็งแกร่งและมีประสิทธิภาพด้านการรักษาความมั่นคงปลอดภัยไซเบอร์ภายในองค์กร โดยใช้ RapidMiner Studio ในการวิเคราะห์ข้อมูลการบุกรุก ด้วยอัลกอริทึม 4 แบบ ได้แก่ Decision Tree, Naïve Bayes, Random Forest และ Gradient Boosted Trees ซึ่งใช้งานร่วมกับชุดข้อมูลการบุกรุกทางไซเบอร์ของกองทัพอากาศ เพื่อทำการประเมินและเปรียบเทียบความแม่นของแต่ละอัลกอริทึม โดยใช้การประเมินประสิทธิภาพจากค่าความแม่น และค่าความเที่ยง จากนั้นนำผลการประเมินประสิทธิภาพอัลกอริทึมที่ดีที่สุดจาก 4 แบบ มาเพียงสองอันดับแรกไปสร้างเป็นอัลกอริทึมใหม่ด้วยเทคนิคการเรียนรู้ของเครื่องแบบรวมกลุ่ม ต่อมาจะทำการประเมินและเปรียบเทียบความแม่นของแต่ละอัลกอริทึม ทั้งนี้ ได้ทำการวัดค่าความแม่นและค่าความเที่ยง สุดท้ายแล้วจึงนำอัลกอริทึมที่ผ่านการประเมินประสิทธิภาพสูงสุดไปสร้างเป็นระบบตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่อง ผลการวิจัยพบว่า ระบบตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่อง สำหรับเสริมสร้างความแข็งแกร่งและมีประสิทธิภาพด้านการรักษาความมั่นคงปลอดภัยไซเบอร์ในองค์กร ที่ดำเนินการตามกระบวนการพัฒนาซอฟต์แวร์ร่วมกับ การสร้างระบบตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่องแบบรวมกลุ่มด้วยอัลกอริทึม Gradient Boosted Trees ที่ทำงานร่วมกับอัลกอริทึม Naïve Bayes โดยใช้เทคนิค Stacking นั้น สามารถให้ผลการทำนาย มีค่าความแม่นเท่ากับร้อยละ 99.77 และค่าความเที่ยงเท่ากับร้อยละ 88.59

คำสำคัญ **:** การรักษาความมั่นคงปลอดภัยไซเบอร์ การบุกรุกทางไซเบอร์ การเรียนรู้ของเครื่อง

Journal of Engineering and Digital Technology (JEDT)
Vol.12 No.2 July - December 2024

Real-time Intrusion Detection System Using Machine Learning for Enhanced and Efficient Cybersecurity in Organizations

Natdanai Kongkhunthod[1*] Prasong Praneetpolgrang[2] Payap Sirinam[3]

[1*2,3]*Master of Engineering Program in Defense Technology, Office of Graduate Studies, Navaminda Kasatriyadhiraj Royal Thai Air Force Academy, Bangkok, Thailand*

*Corresponding Author. E-mail address: natdanai_k@rtaf.mi.th

Received: 13 April 2024; Revised: 12 May 2024; Accepted: 28 May 2024
Published online: 26 December 2024

Abstract

The purpose of this research is to create a real-time cyber intrusion detection system using machine learning to strengthen and enhance cybersecurity within organizations. RapidMiner Studio was used to analyze intrusion data with 4 algorithms: Decision Tree, Naïve Bayes, Random Forest, and Gradient Boosted Trees. The algorithms were evaluated and compared using the Royal Air Force Cyber Intrusion Detection Dataset (RTAF Dataset) to determine their accuracy based on precision and recall values. The top 2 from 4 performing algorithms were then combined using ensemble machine learning techniques. The resulting algorithm was evaluated and compared based on precision and recall values. The algorithm with the highest performance was then used to create a real-time cyber intrusion detection system using machine learning. The results showed that the real-time cyber intrusion detection system using machine learning for strengthening and enhancing cybersecurity in organizations, developed using a software development process combined with ensemble machine learning-based real-time cyber intrusion detection using the Gradient Boosted Trees algorithm in conjunction with the Naïve Bayes algorithm using the Stacking technique, can provide prediction results with an accuracy of 99.77% and a precision of 88.59%.

***Keywords*:** Cybersecurity, Cyber intrusion, Machine learning

1) บทนำ

ในปัจจุบันเทคโนโลยีดิจิทัลได้มีบทบาทที่สำคัญเป็นอย่างมากต่อการพลิกฟื้น ปรับปรุงและพัฒนายกระดับประสิทธิภาพการทำงานขององค์กรหรือหน่วยงาน อีกทั้ง เทคโนโลยีดิจิทัลอุบัติใหม่มีการพัฒนาก้าวหน้าอย่างรวดเร็วจนเป็นเทคโนโลยีดิสรัปชันในขณะเดียวกัน โลกต้องเผชิญกับความเสี่ยงทางดิจิทัลในรูปแบบใหม่ ๆ นั่นคือ ภัยคุกคามทางดิจิทัลหรือไซเบอร์ เช่น มัลแวร์ (malware) และการหลอกลวงทางไซเบอร์ หากแต่ภัยคุกคามดังกล่าวได้พัฒนาศักยภาพของตัวเองให้มีการโจมตีที่ซับซ้อนมากยิ่งขึ้น ซึ่งส่งผลกระทบต่อการรักษาความมั่นคงปลอดภัยไซเบอร์ ความมั่นคงทางเศรษฐกิจ และความมั่นคงของชาติ ในระดับที่มาก

อย่างไรก็ตาม องค์กรต่าง ๆ ได้ตอบสนองต่อภัยคุกคามทางไซเบอร์ที่เกิดขึ้นค่อนข้างล่าช้าอันเนื่องมาจากการขาดเทคโนโลยีและทักษะในการรับมือกับภัยคุกคามทางไซเบอร์ที่เหมาะสม ดังนั้น การเตรียมความพร้อมในการรับมือต่อผลกระทบที่จะเกิดขึ้นภายหลังการถูกโจมตีทางไซเบอร์ รวมทั้งการสร้างความตระหนักรู้ทางไซเบอร์ให้กับบุคลากรในองค์กร จึงมีความสำคัญและจำเป็นอย่างยิ่ง

สำนักงานคณะกรรมการการรักษาความมั่นคงปลอดภัยไซเบอร์แห่งชาติ (สกมช.) เป็นหน่วยงานที่มีหน้าที่เกี่ยวกับการรักษาความมั่นคงปลอดภัยไซเบอร์ในประเทศไทย ได้ระบุอย่างชัดเจนจากข้อมูลสถิติและสถานการณ์ทางไซเบอร์ของไทยในปี พ.ศ. 2565 ที่ผ่านมาพบว่า หน่วยงานภาครัฐถูกโจมตีมากที่สุด รองลงมาคือหน่วยงานเอกชน ทั้งนี้ การโจมตีดังกล่าวได้ส่งผลกระทบไปถึงความมั่งคงปลอดภัยของข้อมูล และความมั่นคงของประชาชนในประเทศ อีกทั้ง จากสถิติดังกล่าว ยังพบว่าปัญหาหลักในด้านช่องโหว่ของการรักษาความมั่นคงปลอดภัยไซเบอร์ (malicious code) มาเป็นอันดับ 1 ของช่องทางการโจมตี คิดเป็นร้อยละ 54 ของการโจมตีทั้งหมด ในขณะที่ภัยคุกคามในรูปแบบของช่องโหว่ในโปรแกรม (vulnerability) ตามมาเป็นอันดับ 2 คิดเป็นร้อยละ 24 และความพยายามบุกรุก (intrusion attempts) คิดเป็นร้อยละ 11 ตามลำดับ [1]

กองทัพอากาศ เป็นหน่วยงานของภาครัฐที่มีบทบาทสำคัญในการรักษาความมั่งคงและพัฒนาประเทศ ดังจะเห็นได้จากนโยบายกองทัพอากาศดิจิทัลที่ได้ดำเนินการมาตั้งแต่ปี พ.ศ. 2558 อนึ่ง ยุทธศาสตร์กองทัพอากาศ 20 ปี (พ.ศ. 2561–2580) ยังได้กล่าวถึงภัยคุกคามในมิติไซเบอร์ (cyber domain) และให้ความสำคัญกับการป้องกันระบบเทคโนโลยีสารสนเทศและการสื่อสาร รวมทั้งระบบเครือข่ายคอมพิวเตอร์ของกองทัพอากาศ ที่อาจตกเป็นเป้าหมายการโจมตีทางไซเบอร์ในลักษณะต่าง ๆ ได้แก่ การเจาะระบบคอมพิวเตอร์ (hacking) การสอดแนมข้อมูลคอมพิวเตอร์ (spyware) การดักรับข้อมูลคอมพิวเตอร์ (sniffing) และการโจมตีที่ปฏิเสธการให้บริการ (Denial of Service: DoS) [2] เป็นต้น ซึ่งเป็นการโจมตีเพื่อขโมยคัดลอก ดัดแปลง ใช้ประโยชน์ข้อมูลหรือเพื่อขู่กรรโชกทรัพย์ รวมทั้งเพื่อควบคุมการทำงานของโครงสร้างพื้นฐานสำคัญ

เมื่อพิจารณาถึงความเสี่ยงที่อาจจะเกิดขึ้นและมีผลกระทบต่อกองทัพอากาศ โดยเฉพาะต้องเผชิญหน้ากับการโจมตีทางไซเบอร์แล้ว การพัฒนาระบบป้องกันภัยคุกคามทางไซเบอร์ด้วยปัญญาประดิษฐ์และเทคนิคการเรียนรู้ของเครื่องผ่านโปรแกรม RapidMiner Studio จึงเป็นเทคนิคที่น่าสนใจในการป้องกันและตอบสนองต่อการโจมตีทางไซเบอร์ ทั้งนี้ จะทำให้เรามั่นใจได้ว่ากองทัพอากาศสามารถเปลี่ยนผ่านไปสู่การเป็นกองทัพอากาศดิจิทัลที่สมาร์ตและมีความปลอดภัยจากภัยคุกคามทางไซเบอร์ที่อาจเกิดขึ้นได้ในอนาคต

จากที่กล่าวมาข้างต้น ผู้วิจัยจึงมีความสนใจที่จะพัฒนาระบบตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ ด้วยการเรียนรู้ของเครื่อง เพื่อทำการตรวจจับการบุกรุกทางไซเบอร์ ที่ใช้โปรแกรม RapidMiner Studio ซึ่งเป็นการออกแบบและควบคุมการทำงานของระบบตรวจจับการบุกรุกทางไซเบอร์ให้เป็นอัตโนมัติ

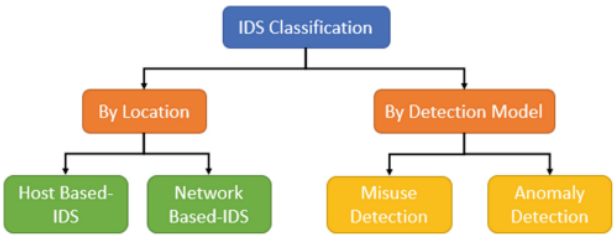
2) วัตถุประสงค์ของงานวิจัย

เพื่อสร้างระบบตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่อง สำหรับเสริมสร้างความแข็งแกร่งและมีประสิทธิภาพด้านการรักษาความมั่นคงปลอดภัยไซเบอร์ในองค์กร

3) ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

3.1) ระบบตรวจจับการบุกรุกทางไซเบอร์ (Intrusion Detection Systems: IDS)

เป็นระบบที่ออกแบบมาเพื่อตรวจจับและแจ้งเตือนเมื่อมีการพยายามบุกรุกในเครือข่ายคอมพิวเตอร์ เป้าหมายหลักของ IDS คือการปกป้องข้อมูลและทรัพยากรทางไซเบอร์ [3]–[6] ลักษณะการทำงานของระบบคือการตรวจสอบและวิเคราะห์กิจกรรมต่าง ๆ ในเครือข่ายเพื่อระบุพฤติกรรมที่ไม่ปกติ ซึ่งอาจรวมถึงการพยายามบุกรุกการโจมตีทางไซเบอร์หรือการกระทำที่ผิดปกติอื่น ๆ ระบบ IDS มีหลายประเภท แต่ละประเภทมีวิธีการทำงานที่แตกต่างกัน แสดงดังรูปที่ 1



รูปที่ 1 : ประเภทของ Intrusion Detection Systems [7]

จากรูปที่ 1 อธิบายประเภทและคุณสมบัติของระบบตรวจจับการบุกรุกทางไซเบอร์ (IDS) ดังนี้

3.1.1) ระบบตรวจจับการบุกรุกแบบระบุลักษณะ (Signature-Based IDS) เป็นการเปรียบเทียบเหตุการณ์ที่เข้ามาในเครือข่ายกับเหตุการณ์ที่เป็นการบุกรุกซึ่งจัดเก็บไว้ในฐานข้อมูล เมื่อเปรียบเทียบแล้วตรงกันแสดงว่าเป็นเหตุการณ์บุกรุก

3.1.2) ระบบตรวจจับการบุกรุกแบบระบุเหตุการณ์ผิดปกติ (Anomaly-Based IDS) เป็นการเปรียบเทียบเหตุการณ์ที่เข้ามาในเครือข่ายกับเหตุการณ์ที่ปกติซึ่งจัดเก็บไว้ในฐานข้อมูล หากเปรียบเทียบแล้วไม่ตรงกันแสดงว่าเหตุการณ์นั้นเป็นการบุกรุก

3.1.3) ระบบตรวจจับการบุกรุกแบบระบุการบุกรุกบนเครือข่าย (Network-Based IDS) ระบบจะทำงานในเครือด้วยการติดตามตรวจสอบสัญญาณจราจรที่อยู่ในรูปของแพ็กเก็ตข้อมูลเพื่อตรวจสอบว่าแพ็กเก็ตใดมีลักษณะที่ผิดปกติ

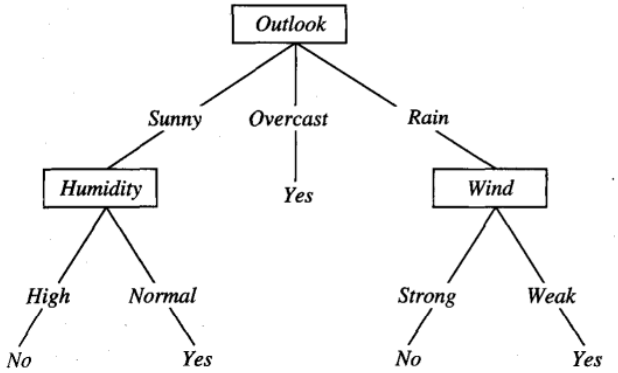
3.1.4) ระบบตรวจจับการบุกรุกบนเครื่องคอมพิวเตอร์แม่ข่าย (Host-based IDSs) ระบบจะตรวจสอบสัญญาณจราจรที่อยู่ในรูปของแพ็กเก็ตข้อมูลและตรวจสอบการใช้งานโปรแกรมประยุกต์ในเครื่องคอมพิวเตอร์แม่ข่ายจากแฟ้มลงบันทึกเข้าออกเพื่อตรวจสอบว่าแพ็กเก็ตใดมีลักษณะที่ผิดปกติ

3.2) เครื่องมือที่ใช้ในการวิเคราะห์ข้อมูลการบุกรุกทางไซเบอร์

โปรแกรม RapidMiner Studio เป็นเครื่องมือสำหรับการวิเคราะห์เชิงคาดการณ์และการทำเหมืองข้อมูล [8] มีความสามารถในการจัดการข้อมูลที่มีขนาดใหญ่ และสร้างอัลกอริทึมในรูปแบบตามที่กำหนด รวมถึงการแสดงผลการทดสอบประสิทธิภาพได้

3.3) อัลกอริทึมที่ใช้จำแนกภัยคุกคามทางทางไซเบอร์

3.3.1) อัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) เป็นเทคนิคการเรียนรู้ของเครื่องที่ใช้ระบุประเภทภัยคุกคามทางไซเบอร์ (classification) และการถดถอย (regression) ที่สร้างโครงสร้างเป็นต้นไม้ แต่ละโหนดของต้นไม้จะตัดสินใจตามเงื่อนไขของคุณลักษณะ (features) และแต่ละใบ (leaf) จะเก็บค่าการตัดสินใจหรือการทำนาย ข้อดีของ Decision Tree คือความสามารถในการนำเสนอความรู้ในรูปแบบที่อ่านและเข้าใจได้ [4], [9] แสดงดังรูปที่ 2

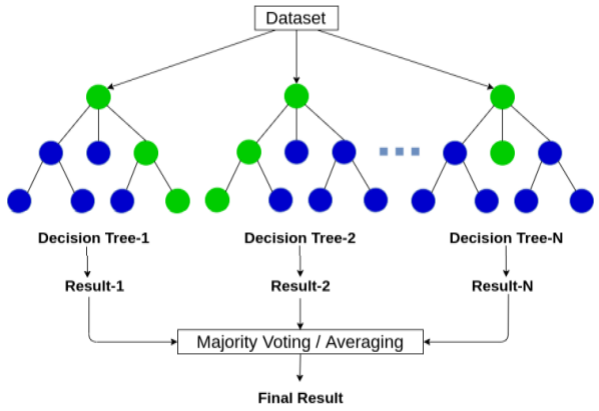


รูปที่ 2 : การทำงานของ Decision Tree [10]

3.3.2) อัลกอริทึมนาอีฟเบย์ (Naïve Bayes) เป็นเทคนิคการเรียนรู้ของเครื่องที่ใช้อัลกอริทึมในการจำแนกข้อมูลที่ใช้ความน่าจะเป็น และวิธีการทางสถิติจากกฎของเบย์ เพื่อหาความน่าจะเป็นที่ถูกต้องที่สุด จึงจำเป็นต้องอาศัยข้อมูลก่อนหน้ากับข้อมูลปัจจุบัน เพื่อหาสมมติฐานที่ดีที่สุด [9], [11], [12] ดังสมการที่ (1)

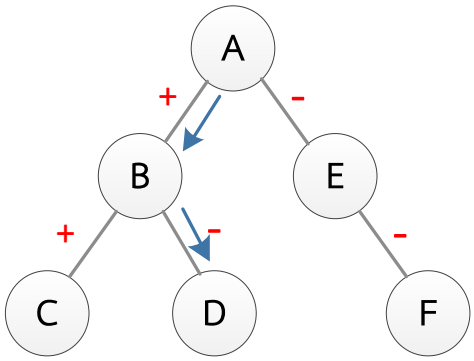
$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \quad (1)$$

3.3.3) อัลกอริทึมการตัดสินใจแบบสุ่ม (Random Forest) เป็นเทคนิคการเรียนรู้ของเครื่องที่ประกอบด้วยการสร้างกลุ่มของต้นไม้ตัดสินใจจำนวนมาก โดยใช้ชุดข้อมูลย่อยที่ถูกสุ่มมาจากชุดข้อมูลต้นฉบับ และใช้การโหวตเพื่อทำนายผลลัพธ์ ซึ่งการใช้ชุดข้อมูลย่อยและคุณลักษณะที่ถูกสุ่มจะทำให้อัลกอริทึมมีความแข็งแกร่งต่อการฝึกฝนที่เกินความจำเป็น (overfitting) [9] แสดงดังรูปที่ 3



รูปที่ 3 : การทำงานของ Random Forest [13]

3.3.4) อัลกอริทึม Gradient Boosted Trees เป็นเทคนิคการเรียนรู้ของเครื่องที่รวมกลุ่มของต้นไม้ตัดสินใจ ด้วยวิธีการเพิ่มต้นไม้แต่ละต้นแบบต่อเนื่องในการฝึกฝนเพื่อปรับปรุงความผิดพลาดจากต้นไม้ก่อนหน้านี้ และทำการประเมินผลของต้นไม้แต่ละต้นจนกว่าจะได้ต้นไม้ตัดสินใจที่สมบูรณ์ที่สุด [9], [14] ดังรูปที่ 4



รูปที่ 4 : การทำงานของ Gradient Boosted Trees [14]

3.3.5) เทคนิคการเรียนรู้ของเครื่องแบบรวมกลุ่ม (Ensemble Method) วิธีการรวมอัลกอริทึมด้วยกระบวนการการเรียนรู้ของเครื่องที่ใช้หลายโมเดล เพื่อทำนายผลลัพธ์ในการจำแนกประเภทหรือการถดถอย เพื่อเพิ่มประสิทธิภาพและความแม่น (accuracy) ในการทำนาย นอกจากนั้น ยังมีเทคนิคที่ได้รับความนิยมในการประยุกต์ใช้งาน [9], [15]–[18] ดังนี้

1) เทคนิค Bagging หรือ Bootstrap Aggregating คือ การสร้างโมเดลหลาย ๆ ตัวอย่างอิสระจากชุดข้อมูลย่อยที่สุ่มมาจากชุดข้อมูลฝึกฝน แต่ละโมเดลจะทำนายอย่างอิสระและผลที่ได้มักจะถูก “โหวต” หรือเฉลี่ยเพื่อให้ได้ผลลัพธ์สุดท้าย

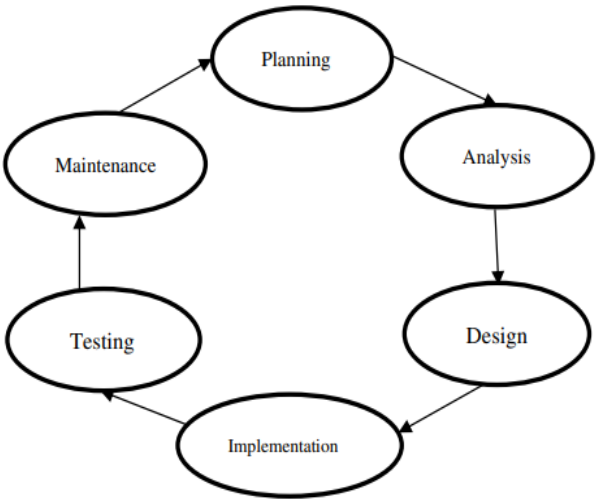
2) เทคนิค Boosting คือ การสร้าง และรวมโมเดลหลาย ๆ ตัว เพื่อเพิ่มประสิทธิภาพในการทำนายหรือการจำแนกประเภท โดยมุ่งเน้นในการสร้างโมเดลทีละตัว ทั้งนี้ แต่ละโมเดลจะพยายามแก้ไขข้อผิดพลาดจากโมเดลก่อนหน้า

3) เทคนิค โหวต ลงมติ หรือเลือก (Voting) คือ กระบวนการในการทำนายผลลัพธ์ของข้อมูลที่อาศัยโมเดลหลายตัว และใช้ผลลัพธ์ที่ได้จากโมเดลเหล่านั้นมาใช้เป็นเกณฑ์สำหรับการทำนายผลลัพธ์สุดท้าย วิธีนี้แบ่งได้เป็น 2 ประเภท คือ การโหวตที่ใช้โมเดลมากที่สุดเป็นเกณฑ์ (Hard Voting) และการโหวตจะพิจารณาความน่าจะเป็นของแต่ละโมเดล (Soft Voting)

4) เทคนิค Stacked Generalization (Stacking) คือ กระบวนการที่นำเอาผลลัพธ์จากโมเดลหลายตัวมาเป็นอินพุตของโมเดลที่สอง เรียกว่า Meta-Model เพื่อทำนายผลลัพธ์สุดท้าย วิธีนี้ช่วยให้สามารถนำจุดเด่นของโมเดลหลายตัวมารวมกันเพื่อให้ได้โมเดลที่มีประสิทธิภาพมากขึ้น

3.4) กระบวนการพัฒนาซอฟต์แวร์ (Software Development Life Cycle: SDLC)

เป็นกระบวนการที่ครอบคลุมทุกขั้นตอนของการพัฒนาซอฟต์แวร์ ตั้งแต่ การวางแผน การวิเคราะห์ความต้องการ การออกแบบ การพัฒนา การทดสอบ และการบำรุงรักษา [19] แสดงดังรูปที่ 5

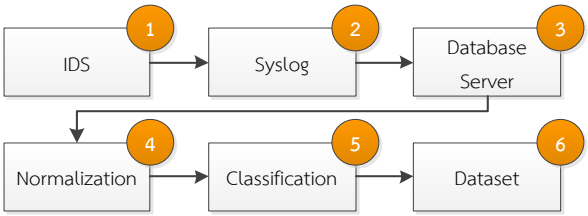


รูปที่ 5 : กระบวนการพัฒนาซอฟต์แวร์ [19]

4) วิธีดำเนินการวิจัย

4.1) ชุดข้อมูลที่ใช้ในการวิจัย (Dataset)

ในการวิจัยในครั้งนี้ผู้วิจัยได้เก็บรวบรวมข้อมูลจากระบบตรวจจับการบุกรุกทางไซเบอร์ของศูนย์ไซเบอร์กองทัพอากาศ เพื่อจัดทำเป็นชุดข้อมูล [3] สำหรับนำไปฝึกสอนโมเดล และได้ทำการจัดเตรียมชุดข้อมูล มีขั้นตอนดังนี้ 1) ดักจับข้อมูลการจราจรทางคอมพิวเตอร์จากระบบตรวจจับการบุกรุกทางไซเบอร์ (IDS) 2) บันทึกข้อมูลสู่ระบบคอมพิวเตอร์ในรูปแบบของ System Log (Syslog) 3) บันทึกข้อมูลเข้าสู่ระบบฐานข้อมูล (database server) 4) แปลงข้อมูลจากรูปแบบ System Log ให้อยู่ในรูปแบบที่มีโครงสร้าง (normalization) 5) จำแนกประเภทภัยคุกทางไซเบอร์ (classification) 6) ส่งข้อมูลออกเป็นชุดข้อมูล [15] แสดงดังรูปที่ 6



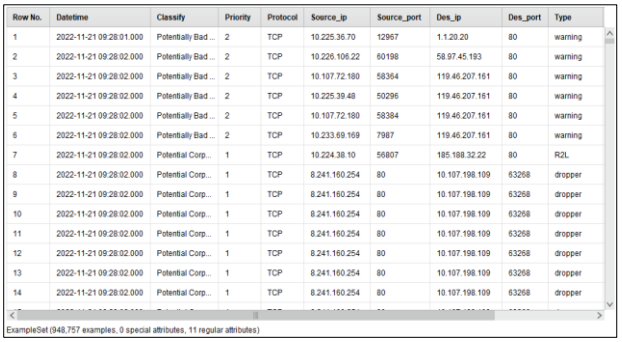
รูปที่ 6 : การจัดเตรียมชุดข้อมูล (dataset)

จากรูปที่ 6 ทำให้ผู้วิจัยได้ชุดข้อมูลการตรวจจับการบุกรุกทางไซเบอร์ของกองทัพอากาศ (RTAF Dataset) จำนวน 9 แอตทริบิวต์ แสดงดังตารางที่ 1 และข้อมูลอินพุต ดังรูปที่ 7

มีข้อมูลการตรวจจับการบุกรุกทางไซเบอร์ของกองทัพอากาศ จำนวน 948,757 รายการ แบ่งตามประเภทภัยคุกคามทางไซเบอร์ ดังตารางที่ 2

ตารางที่ 1 : ชื่อแอตทริบิวต์

ลำดับ	ชื่อแอตทริบิวต์	ประเภท
1	Classify	Discrete
2	Datetime	Discrete
3	Des_ip	Discrete
4	Des_port	Discrete
5	Priority	Discrete
6	Protocol	Discrete
7	Source_ip	Discrete
8	Source_port	Discrete
9	Type	Discrete



Row No.	Datetime	Classify	Priority	Protocol	Source_ip	Source_port	Des_ip	Des_port	Type
1	2022-11-21 09:28:01.000	Potentially Bad ...	2	TCP	10.225.36.70	12967	1.1.20.20	80	warning
2	2022-11-21 09:28:02.000	Potentially Bad ...	2	TCP	10.226.106.22	60198	58.97.45.193	80	warning
3	2022-11-21 09:28:02.000	Potentially Bad ...	2	TCP	10.107.72.180	58364	119.46.207.161	80	warning
4	2022-11-21 09:28:02.000	Potentially Bad ...	2	TCP	10.225.39.48	50296	119.46.207.161	80	warning
5	2022-11-21 09:28:02.000	Potentially Bad ...	2	TCP	10.107.72.180	58384	119.46.207.161	80	warning
6	2022-11-21 09:28:02.000	Potentially Bad ...	2	TCP	10.233.69.169	7987	119.46.207.161	80	warning
7	2022-11-21 09:28:02.000	Potential Corp...	1	TCP	10.224.38.10	56807	185.188.32.22	80	R2L
8	2022-11-21 09:28:02.000	Potential Corp...	1	TCP	8.241.160.254	80	10.107.198.109	63268	dropper
9	2022-11-21 09:28:02.000	Potential Corp...	1	TCP	8.241.160.254	80	10.107.198.109	63268	dropper
10	2022-11-21 09:28:02.000	Potential Corp...	1	TCP	8.241.160.254	80	10.107.198.109	63268	dropper
11	2022-11-21 09:28:02.000	Potential Corp...	1	TCP	8.241.160.254	80	10.107.198.109	63268	dropper
12	2022-11-21 09:28:02.000	Potential Corp...	1	TCP	8.241.160.254	80	10.107.198.109	63268	dropper
13	2022-11-21 09:28:02.000	Potential Corp...	1	TCP	8.241.160.254	80	10.107.198.109	63268	dropper
14	2022-11-21 09:28:02.000	Potential Corp...	1	TCP	8.241.160.254	80	10.107.198.109	63268	dropper

ExampleSet (948,757 examples, 0 special attributes, 11 regular attributes)

รูปที่ 7 : ข้อมูลอินพุต

ตารางที่ 2 : ประเภทและจำนวนภัยคุกคามทางไซเบอร์

ประเภทภัยคุกคามทางไซเบอร์	จำนวน	อัตราส่วน
Normal	5,319	0.5606%
Bitcoin	5,782	0.6094%
botnet	484	0.0510%
compromise	171	0.0180%
data_leak	515	0.0543%
dropper	53,529	5.6420%
Injection	14,831	1.5632%
phishing_Email	2,613	0.2754%
phishing_web	2,268	0.2390%
probe	690	0.0727%
R2L	96,175	10.1369%
trojan	622	0.0656%
U2R	2,359	0.2486%
warning	676,254	71.2779%
Web_Malware	87,124	9.1830%
worm	12	0.0013%
XSS	9	0.0009%
รวม	948,757	100%

4.2) การเตรียมข้อมูล (Data Preparation)

ผู้วิจัยได้ทำการเตรียมข้อมูลให้เหมาะสมและการคัดเลือกคุณลักษณะข้อมูล (feature selection) ที่มีความครอบคลุมพร้อมจัดให้อยู่ในรูปแบบที่นำไปใช้งานร่วมกับโปรแกรม RapidMiner Studio สำหรับการกำหนดคุณลักษณะผลการทำนาย ในที่นี้ได้กำหนดให้ประเภทภัยคุกคามทางไซเบอร์ ในชื่อแอตทริบิวต์ Type เป็นฉลากหรือป้าย (label) จากนั้นทำการแบ่งกลุ่มข้อมูลเพื่อทดสอบ (cross-validation) ด้วยวิธี 10-fold cross-validation จากนั้นแบ่งข้อมูลออกเป็น 2 ส่วนคือ ส่วนแรกใช้สำหรับการฝึกสอนเพื่อให้อัลกอริทึมเกิดการเรียนรู้ (train set) และส่วนที่

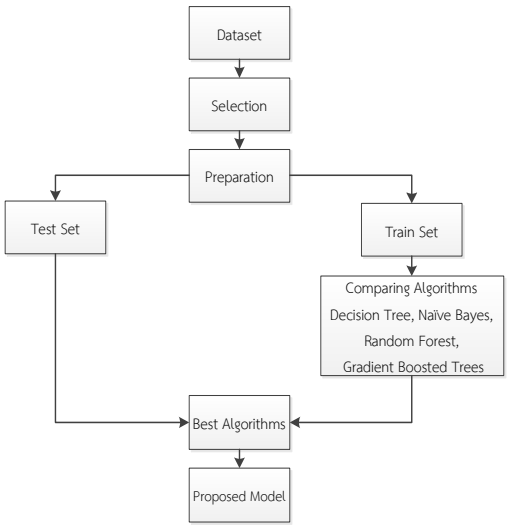
สองใช้สำหรับการทดสอบอัลกอริทึม (test set) จากนั้นทำการทดสอบโดยการเปลี่ยนชุดข้อมูลทดสอบ ตั้งแต่ส่วนที่ 1 เป็นชุดทดสอบและส่วนที่ 2 ถึง 9 เป็นชุดสำหรับการเรียนรู้ไปจนถึงชุดทดสอบส่วนที่ 10 และส่วนที่ 1 ถึง 9 เป็นชุดสำหรับการเรียนรู้ เมื่อทำจนครบ 10 ครั้ง และนำค่าความแม่นของการพยากรณ์ที่ได้ทั้งหมดมาหาค่าเฉลี่ย [20]

4.3) การสร้างโมเดลตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่อง

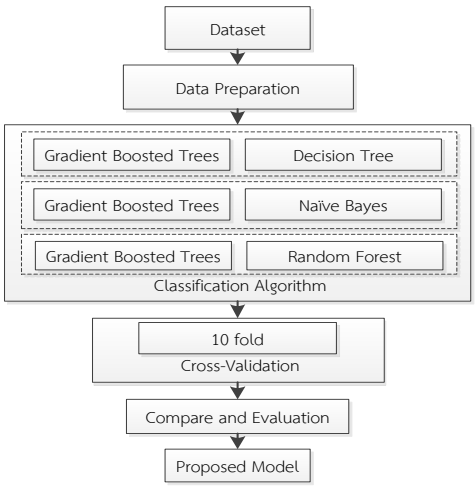
งานวิจัยนี้ ผู้วิจัยได้สร้างโมเดลตรวจจับภัยคุกคามทางไซเบอร์ [21]–[27] ในกองทัพอากาศด้วยการเรียนรู้ของเครื่อง กับเทคนิคการเรียนรู้ของเครื่อง ผู้วิจัยได้ทบทวนวรรณกรรมที่เกี่ยวข้องกับการสร้างโมเดลตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่องแบบรวมกลุ่ม โดยเลือกเทคนิคการเรียนรู้แบบกลุ่มแบบโหวต ลงมติ หรือเลือก (voting) และ Stacked Generalization (stacking) มาประยุกต์ใช้ จากนั้นทำการประเมินและเปรียบเทียบความแม่นของแต่ละอัลกอริทึมประเภทต่าง ๆ จากนั้นนำผลที่ได้จากการวิเคราะห์ดังกล่าวไปสร้างระบบตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่องตามขั้นตอน สามารถอธิบายขั้นตอนได้ดังนี้

การสร้างโมเดล (model building) ผู้วิจัยได้สร้างโมเดลการเรียนรู้ของเครื่องแบบมีผู้สอนด้วยโปรแกรม RapidMiner Studio เป็นการใช้อัลกอริทึมการจำแนกข้อมูลการบุกรุก 4 อัลกอริทึม ประกอบด้วยอัลกอริทึม Decision Tree, Naïve Bayes, Random Forest และ Gradient Boosted Trees กับชุดข้อมูลการตรวจจับการบุกรุกทางไซเบอร์ของกองทัพอากาศ (RTAF Dataset) เพื่อเปรียบเทียบความแม่นของแต่ละอัลกอริทึมด้วยสมการวัดค่าความแม่น (accuracy) ค่าความเที่ยง (precision) ค่าความระลึก (recall) และค่าความถ่วงดุล (F1-score) [28] แสดงดังรูปที่ 8

จากนั้น นำผลการประเมินและเปรียบเทียบความแม่นของอัลกอริทึม เพื่อหาอัลกอริทึมการจำแนกภัยคุกตามทางไซเบอร์ที่ดีที่สุดไปสร้างเป็นโมนเดลใหม่ ด้วยเทคนิคการเรียนรู้ของเครื่องแบบรวมกลุ่ม [18] ผลการประเมินและเปรียบเทียบความแม่นของแต่ละอัลกอริทึม พบว่า อัลกอริทึม Gradient Boosted Trees ค่าความแม่น และค่าความเที่ยงที่ดีที่สุด ผู้วิจัยจึงนำอัลกอริทึมดังกล่าวไปใช้เป็นอัลกอริทึมหลักเพื่อสร้างเป็นโมเดลใหม่ด้วยเทคนิคการเรียนรู้ของเครื่องแบบรวมกลุ่มกับ 3 อัลกอริทึมที่มีอันดับรองลงมา เพื่อประเมินและเปรียบเทียบประสิทธิภาพโมเดลจากสมการวัดค่าความแม่น และค่าความเที่ยงที่ดีที่สุด แสดงดังรูปที่ 9



รูปที่ 8 : แผนภูมิการสร้างโมเดลตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่อง



รูปที่ 9 : แผนภูมิการสร้างโมเดลตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่องแบบรวมกลุ่ม

4.4) การประเมินประสิทธิภาพของโมเดล (Model Evaluation)

ผู้วิจัยได้ทำการประเมินประสิทธิภาพโมเดล ด้วยอัลกอริทึมการจำแนกภัยคุกคามทางไซเบอร์ทั้ง 4 อัลกอริทึม ประกอบด้วยอัลกอริทึม Decision Tree, Naïve Bayes, Random Forest และ Gradient Boosted Trees และประเมินประสิทธิภาพโมเดลตรวจจับภัยคุกคามทางไซเบอร์ด้วยเทคนิคการเรียนรู้ของเครื่องแบบรวมกลุ่ม กับชุดข้อมูลการตรวจจับการบุกรุกทางไซเบอร์ของกองทัพอากาศ (RTAF Dataset) ที่ใช้สมการวัดค่าความแม่น

ค่าความเที่ยง ค่าความระลึก และค่าความถ่วงดุล [29], [30] ตามสมการที่ (2)–(5) ดังนี้

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \quad (2)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4)$$

$$\text{F1-Score} = 2\frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

โดยที่

True Positive (TP) คือ ข้อมูลที่ทำนายถูกต้อง เมื่อเทียบกับเฉลย

False Positive (FP) คือ ข้อมูลที่ทำนายแล้วไม่ถูกต้อง เมื่อเทียบกับเฉลย

True Negative (TN) คือ ข้อมูลที่อยู่ในเฉลยแต่ไม่มีการทำนาย (ตรงข้ามกับ FN)

Precision คือ ค่าความเที่ยงเกิดจากการนำค่า TP มาเทียบกับ FP

Recall คือ ค่าความระลึกเกิดจากการนำค่า TP มาเทียบกับ FN

F1-Score คือ ค่าเฉลี่ยของ Precision และ Recall

4.5) การออกแบบระบบตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่อง

ผู้วิจัยได้ออกแบบระบบตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่อง สำหรับเสริมสร้างความแข็งแกร่งและมีประสิทธิภาพด้านการรักษาความมั่นคงปลอดภัยไซเบอร์ในองค์กร ตามกระบวนการพัฒนาซอฟต์แวร์ (SDLC) 6 ขั้นตอนดังต่อไปนี้

4.5.1) การวางแผน (Planning) ศึกษา รวบรวม ผลการวิเคราะห์ภัยคุกคาม และอัลกอริทึมที่มีความเหมาะสม

4.5.2) การวิเคราะห์ความต้องการ (Analysis) สอบถามความคิดเห็นจากผู้ปฏิบัติงานในศูนย์ยุทธการทางไซเบอร์ ศูนย์ไซเบอร์กองทัพอากาศ จากนั้นนำมาวิเคราะห์เพื่อทำความเข้าใจ และกำหนดขอบเขต

4.5.3) การออกแบบ (Design) นำผลการวิเคราะห์ความต้องการ มาออกแบบระบบทั้งในส่วนของผู้ใช้งาน (front-end) กับส่วนควบคุมการประมวลผล (back-end) และส่วนเชื่อมต่อกับระบบแลกเปลี่ยนข้อมูล (REST API)

4.5.4) การพัฒนา (Implementation) พัฒนาระบบที่มีการเชื่อมต่อไปยังระบบแลกเปลี่ยนข้อมูล (REST API) จากนั้นนำข้อมูลที่ได้จากระบบดังกล่าวมานำเสนอในรูปแบบของจินตทัศน์ข้อมูล (data visualization)

4.5.5) การทดสอบ (Testing) ทดสอบข้อบกพร่องของระบบตามกระบวนการทดสอบซอฟต์แวร์ (Software Testing Life Cycle: STLC) เพื่อให้แน่ใจว่าระบบสามารถทำงานได้อย่างถูกต้อง

4.5.6) การบำรุงรักษา (Maintenance) บำรุงรักษาอย่างต่อเนื่อง เพื่อแก้ไขปัญหาที่อาจจะเกิดขึ้น ประกอบด้วย การอัปเดตระบบปฏิบัติการ การอัปเดตระบบฐานข้อมูล และเพิ่มประสิทธิภาพของระบบปฏิบัติการ

5) ผลการวิจัย

5.1) ผลการประเมินและเปรียบเทียบประสิทธิภาพความแม่นของแต่ละอัลกอริทึม

ผู้วิจัยได้ใช้ชุดข้อมูลการตรวจจับการบุกรุกทางไซเบอร์ของกองทัพอากาศ ซึ่งเป็นชุดข้อมูลเฉพาะและเป็นชุดข้อมูลที่มาจากข้อมูลจริง นำเข้ากระบวนการสร้างโมเดลตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่องด้วยโปรแกรม RapidMiner Studio ด้วยเทคนิคการเรียนรู้ของเครื่อง และใช้อัลกอริทึมการจำแนกข้อมูลการบุกรุก 4 อัลกอริทึม ประกอบด้วย อัลกอริทึม Decision Tree, Naïve Bayes, Random Forest และ Gradient Boosted Trees เพื่อประเมินและเปรียบเทียบความแม่นของแต่ละอัลกอริทึม โดยใช้สมการวัดค่าความแม่น ค่าความเที่ยง ค่าความระลึก และค่าความถ่วงดุล สามารถแสดงการประเมินและเปรียบเทียบประสิทธิภาพ ดังตารางที่ 3

ตารางที่ 3 : การประเมินและเปรียบเทียบประสิทธิภาพความแม่นของอัลกอริทึมทั้ง 4 อัลกอริทึม

Algorithms	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree	94.98	79.65	58.12	67.20
Naïve Bayes	98.11	72.26	94.28	81.81
Random Forest	94.98	80.06	58.04	67.29
Gradient Boosted Trees	99.07	80.24	71.48	75.61

จากตารางที่ 3 แสดงผลการประเมินผลและเปรียบเทียบประสิทธิภาพของอัลกอริทึมทั้ง 4 อัลกอริทึมสามารถสรุปได้ดังนี้

อัลกอริทึม Gradient Boosted Trees มีค่าเฉลี่ยความแม่นเท่ากับร้อยละ 99.07 ค่าเฉลี่ยความเที่ยง เท่ากับร้อยละ 80.24 ค่าเฉลี่ยความระลึก เท่ากับร้อยละ 71.48 และค่าเฉลี่ยความถ่วงดุลเท่ากับร้อยละ 75.61

อัลกอริทึม Naïve Bayes มีค่าเฉลี่ยความแม่น เท่ากับร้อยละ 98.11 ค่าเฉลี่ยความเที่ยง เท่ากับร้อยละ 72.26 ค่าเฉลี่ยความระลึก เท่ากับร้อยละ 94.28 และค่าเฉลี่ยความถ่วงดุล เท่ากับร้อยละ 81.81

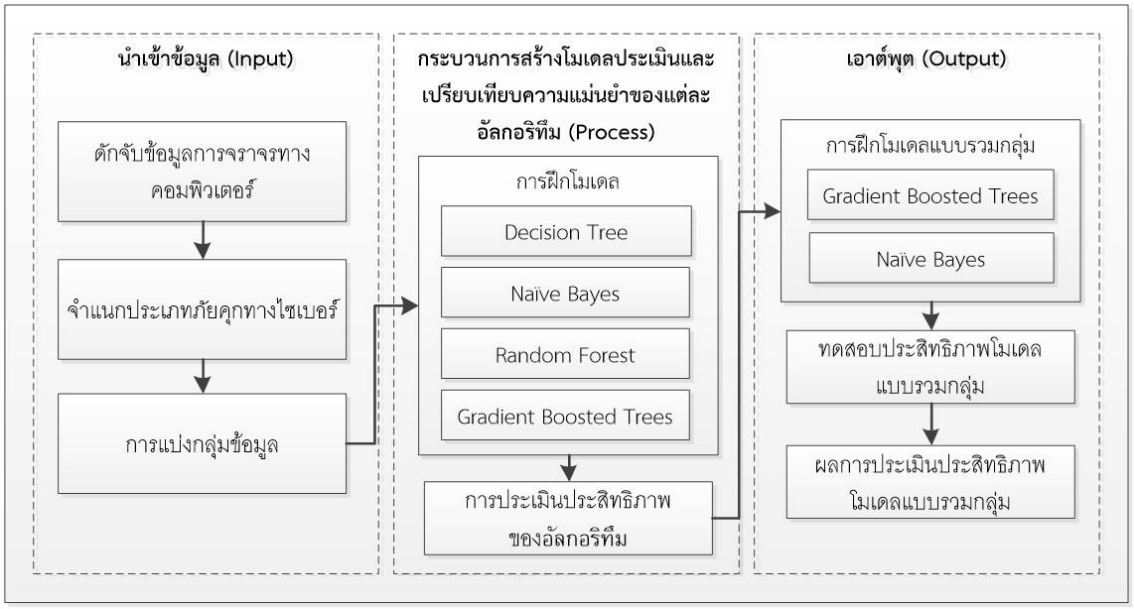
อัลกอริทึม Random Forest มีค่าเฉลี่ยความแม่น เท่ากับร้อยละ 94.98 ค่าเฉลี่ยความเที่ยง เท่ากับร้อยละ 80.06 ค่าเฉลี่ยความระลึก เท่ากับร้อยละ 58.04 และค่าเฉลี่ยความถ่วงดุลเท่ากับร้อยละ 67.29

อัลกอริทึม Decision Tree มีค่าเฉลี่ยความแม่น เท่ากับร้อยละ 90.98 ค่าเฉลี่ยความเที่ยง เท่ากับร้อยละ 79.65 ค่าเฉลี่ยความระลึก เท่ากับร้อยละ 58.12 และค่าเฉลี่ยความถ่วงดุลเท่ากับร้อยละ 67.20 ฉะนั้น อัลกอริทึมที่มีค่าเฉลี่ยความแม่นและค่าเฉลี่ยความเที่ยงสูงที่สุดคือ อัลกอริทึม Gradient Boosted Trees

ผู้วิจัยจึงได้คัดเลือกอัลกอริทึม Gradient Boosted Trees ที่มีค่าความแม่น และค่าความเที่ยงที่สูงสุดไปใช้เป็นอัลกอริทึมหลักสำหรับสร้างระบบตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่อง สำหรับเสริมสร้างความแข็งแกร่งและมีประสิทธิภาพด้านการรักษาความมั่นคงปลอดภัยไซเบอร์ในองค์กร จากนั้นนำผลการการประเมินประสิทธิภาพโมเดล ด้วยอัลกอริทึมการจำแนกภัยคุกตามทางไซเบอร์ที่ดีที่สุด ไปสร้างเป็นโมเดลใหม่ด้วยเทคนิคการเรียนรู้ของเครื่องแบบรวมกลุ่ม [16]–[18] กับอัลกอริทึม 3 อัลกอริทึม ประกอบด้วยอัลกอริทึม Decision Tree, Naïve Bayes และ Random Forest เพื่อประเมินและเปรียบเทียบความแม่นของแต่ละอัลกอริทึม ที่ใช้สมการวัดค่าความแม่น ค่าความเที่ยง ดังรูปที่ 10

จากรูปที่ 10 แสดงการสร้างโมเดลตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่องแบบรวมกลุ่ม ประกอบด้วย 3 ส่วน ดังนี้

ส่วนที่ 1 การนำเข้าข้อมูล (input) ด้วยการดักจับข้อมูลการจราจรทางคอมพิวเตอร์จากเครื่องข่ายคอมพิวเตอร์ในองค์กร จากนั้นทำการจำแนกประเภทภัยคุกคามทางไซเบอร์ จากนั้นทำการแบ่งกลุ่มข้อมูลเพื่อทดสอบด้วยวิธี 10-fold cross-validation และแบ่งชุดข้อมูลออกเป็น 2 ส่วนคือ Train Set และ Test Set ส่วนแรกใช้สำหรับการฝึกสอนเพื่อให้อัลกอริทึมเกิดการเรียนรู้ ส่วนที่สองใช้สำหรับการทดสอบอัลกอริทึม จากนั้นทำการทดสอบโดยการเปลี่ยนชุดข้อมูลทดสอบ ทั้งนี้ ได้เริ่มตั้งแต่ส่วนที่ 1 เป็นชุดทดสอบและส่วนที่ 2 ถึง 9 เป็นชุดสำหรับการเรียนรู้ไปจนถึงชุดทดสอบส่วนที่ 10 และส่วนที่ 1 ถึง 9 เป็นชุดสำหรับการเรียนรู้ เมื่อทำจนครบ 10 ครั้ง และนำค่าความแม่นของการพยากรณ์ที่ได้ทั้งหมดมาหาค่าเฉลี่ย

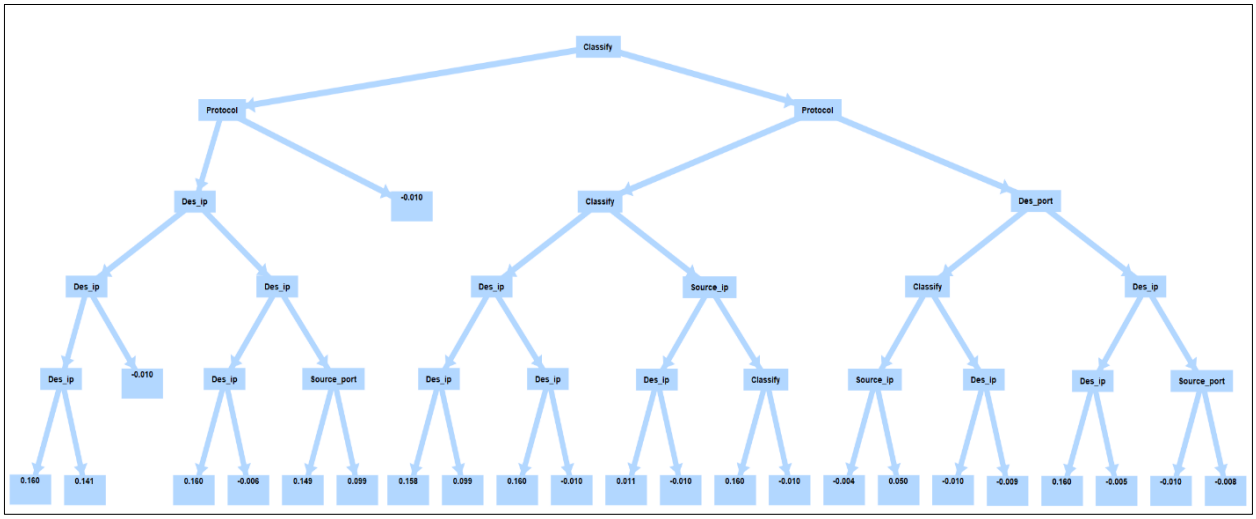


รูปที่ 10 : การสร้างโมเดลตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่องแบบรวมกลุ่ม

ส่วนที่ 2 การสร้างโมเดลเพื่อประเมินประสิทธิภาพของอัลกอริทึม โดยนำข้อมูลที่ได้เตรียมไว้ในส่วนที่ 1 ไปสร้างโมเดล จากนั้นทำการฝึกโมเดลจากนั้นทำการประเมินและเปรียบเทียบประสิทธิภาพ ของอัลกอริทึมทั้ง 4 อัลกอริทึม ประกอบด้วย Decision Tree, Naïve Bayes, Random Forest และ Gradient Boosted Trees กับชุดข้อมูลการตรวจจับการบุกรุกทางไซเบอร์ของกองทัพอากาศ เพื่อเลือกอัลกอริทึมการจำแนกภัยคุกคามทางไซเบอร์ที่ดีที่สุด 2 อันดับไปสร้างเป็นโมเดลแบบรวมกลุ่ม ผลการทดสอบประสิทธิภาพของอัลกอริทึมพบว่า อัลกอริทึม Gradient Boosted Trees มีค่าเฉลี่ยความแม่น ร้อยละ 99.07 และค่าเฉลี่ยความเที่ยง ร้อยละ 80.24 และอัลกอริทึม Naïve Bayes มีค่าเฉลี่ยความแม่น ร้อยละ 98.11 และค่าเฉลี่ยความเที่ยง ร้อยละ 72.26 เป็นเปอร์เซ็นต์สูงสุดจากอัลกอริทึมทั้งหมด ดังตารางที่ 4

Row No.	Datetime	Classify	Priority	Protocol	Source_ip	Source_port	Des_ip	Des_port	prediction(Type)
1	Nov 21, 2022 9:28:02 AM ICT	Potentially Bad Traffic	2	TCP	10.107.72.180	58364	119.46.207.161	80	warning
2	Nov 21, 2022 9:28:02 AM ICT	Potential Corporate Privacy ...	1	TCP	8.241.160.254	80	10.107.198.109	63268	dropper
3	Nov 21, 2022 9:28:02 AM ICT	Potentially Bad Traffic	2	TCP	10.107.184.210	55982	119.46.207.161	80	warning
4	Nov 21, 2022 9:28:02 AM ICT	Potentially Bad Traffic	2	TCP	10.229.6.27	57600	103.235.46.245	80	Injection
5	Nov 21, 2022 9:28:02 AM ICT	Potentially Bad Traffic	2	TCP	10.107.184.210	56082	119.46.207.161	80	warning
6	Nov 21, 2022 9:28:02 AM ICT	Potential Corporate Privacy ...	1	TCP	10.233.73.207	48636	142.250.199.35	80	Web_Malware
7	Nov 21, 2022 9:28:02 AM ICT	Potentially Bad Traffic	2	TCP	10.225.36.70	13000	58.97.45.193	80	warning
8	Nov 21, 2022 9:28:02 AM ICT	Potentially Bad Traffic	2	TCP	10.107.72.180	58554	119.46.207.161	80	warning
9	Nov 21, 2022 9:28:02 AM ICT	Potentially Bad Traffic	2	TCP	10.225.39.48	50486	119.46.207.161	80	warning
10	Nov 21, 2022 9:28:02 AM ICT	Potentially Bad Traffic	2	TCP	10.107.72.180	58403	119.46.207.161	80	warning
11	Nov 21, 2022 9:28:02 AM ICT	Potentially Bad Traffic	2	TCP	10.107.72.180	58413	119.46.207.161	80	warning
12	Nov 21, 2022 9:28:02 AM ICT	Potential Corporate Privacy ...	1	TCP	10.228.255.114	52791	188.172.208.135	80	R2L
13	Nov 21, 2022 9:28:02 AM ICT	Potential Corporate Privacy ...	1	TCP	10.224.38.10	57007	37.252.244.158	80	R2L
14	Nov 21, 2022 9:28:02 AM ICT	Potentially Bad Traffic	2	TCP	10.235.224.75	9993	119.46.126.79	80	warning
15	Nov 21, 2022 9:28:02 AM ICT	Potentially Bad Traffic	2	TCP	10.229.6.27	58080	103.235.46.245	80	Injection
16	Nov 21, 2022 9:28:02 AM ICT	Potentially Bad Traffic	2	TCP	10.235.224.75	9993	119.46.126.79	80	warning
17	Nov 21, 2022 9:28:02 AM ICT	Potentially Bad Traffic	2	TCP	10.233.69.169	8127	119.46.207.161	80	warning
18	Nov 21, 2022 9:28:02 AM ICT	Potential Corporate Privacy ...	1	TCP	10.224.38.10	56879	185.188.32.5	80	R2L
19	Nov 21, 2022 9:28:02 AM ICT	Potential Corporate Privacy ...	1	TCP	10.228.255.114	52791	188.172.208.135	80	R2L
20	Nov 21, 2022 9:28:02 AM ICT	Potential Corporate Privacy ...	1	TCP	10.226.184.162	7804	172.67.75.166	80	R2L

รูปที่ 11 : ผลเอาต์พุต



รูปที่ 12 : ลำดับความสำคัญของแอตทริบิวต์

ตารางที่ 4 : ผลการประเมินและเปรียบเทียบความแม่นของอัลกอริทึมแบบรวมกลุ่ม

Algorithms	Voting				Stacking			
	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Gradient Boosted Trees + Decision Tree	95.04	76.11	52.26	61.97	99.51	81.00	77.67	79.30
Gradient Boosted Trees + Naïve Bayes	98.99	80.03	71.20	75.36	**99.77**	**88.59**	**85.90**	**87.22**
Gradient Boosted Trees + Random Forest	95.06	75.83	52.49	62.04	99.51	81.20	78.31	79.73

ส่วนที่ 3 เอาต์พุต (output) นำอัลกอริทึมที่ผลการประเมินและเปรียบเทียบความแม่นของแต่ละอัลกอริทึมที่มีค่าเฉลี่ยความแม่น และค่าความเที่ยง เป็นเปอร์เซ็นต์สูงสุด 2 อันดับ คืออัลกอริทึม Gradient Boosted Trees และอัลกอริทึม Naïve Bayes ไปสร้างเป็นโมเดลแบบรวมกลุ่มด้วยโปรแกรม RapidMiner Studio ซึ่งเป็นโปรแกรมที่ใช้สำหรับการวิเคราะห์ข้อมูลและการเรียนรู้ของเครื่องที่มีความสามารถในการประมวลผลข้อมูลและสร้างโมเดลทางสถิติโดยไม่ต้องมีทักษะการเขียนโปรแกรม จากนั้นทำการประเมินและเปรียบเทียบความแม่นของแต่ละอัลกอริทึมแบบรวมกลุ่มที่ใช้สมการวัดค่าความแม่น ค่าความเที่ยง ค่าความระลึก และค่าความถ่วงดุล [29], [30] ตามสมการที่ (2)–(5) รวมถึงแสดงผลเอาต์พุตชื่อแอตทริบิวต์ “prediction(Type)” ดังรูปที่ 11 และแสดงลำดับความสำคัญของตัวแปร ดังรูปที่ 12

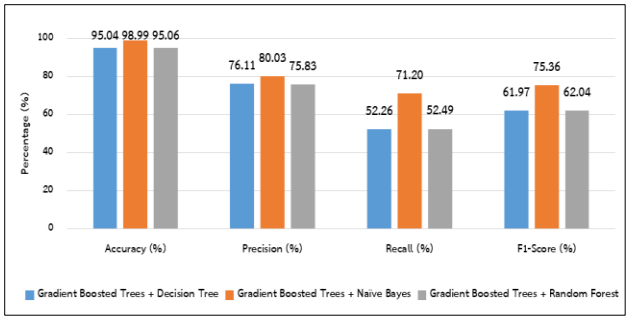
จากรูปที่ 12 แสดงลำดับความสำคัญของแอตทริบิวต์ พบว่าแอตทริบิวต์ Classify มีความสำคัญมากที่สุด รองลงมาคือแอตทริบิวต์ Protocol

จากตารางที่ 4 สรุปผลการประเมินและเปรียบเทียบความแม่นของอัลกอริทึมแบบรวมกลุ่ม ได้ดังนี้ อัลกอริทึม Gradient Boosted Trees ทำงานร่วมกับอัลกอริทึม Decision Tree ใช้เทคนิคการ Voting มีค่าเฉลี่ยความแม่นเท่ากับร้อยละ 95.04 ค่าเฉลี่ยความเที่ยง เท่ากับร้อยละ 76.11 ค่าเฉลี่ยความระลึก เท่ากับร้อยละ 52.26 และค่าเฉลี่ยความถ่วงดุล เท่ากับร้อยละ 61.97 และ Stacking มีค่าเฉลี่ยความแม่น เท่ากับร้อยละ 99.51 ค่าเฉลี่ยความเที่ยง เท่ากับร้อยละ 81.00 ค่าเฉลี่ยความระลึก เท่ากับร้อยละ 77.67 และค่าเฉลี่ยความถ่วงดุล เท่ากับร้อยละ 79.30

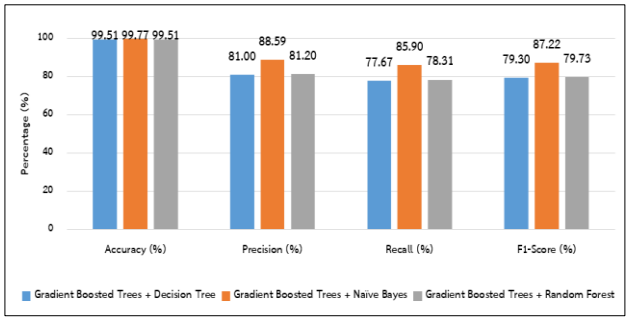
อัลกอริทึม Gradient Boosted Trees ทำงานร่วมกับอัลกอริทึม Naïve Bayes ใช้เทคนิคการ Voting มีค่าเฉลี่ยความแม่น เท่ากับร้อยละ 98.99 ค่าเฉลี่ยความเที่ยง เท่ากับร้อยละ 80.03 ค่าเฉลี่ยความระลึก เท่ากับร้อยละ 71.20 และค่าเฉลี่ยความถ่วงดุล เท่ากับร้อยละ 75.36 และ Stacking มีค่าเฉลี่ยความแม่น เท่ากับร้อยละ 99.77 ค่าเฉลี่ยความเที่ยง เท่ากับร้อยละ 88.59 ค่าเฉลี่ยความระลึก เท่ากับร้อยละ 85.90 และค่าเฉลี่ยความถ่วงดุล เท่ากับร้อยละ 87.22

อัลกอริทึม Gradient Boosted Trees ทำงานร่วมกับอัลกอริทึม Random Forest ใช้เทคนิคการ Voting มีค่าเฉลี่ยความแม่น เท่ากับร้อยละ 95.06 ค่าเฉลี่ยความเที่ยง เท่ากับร้อยละ 75.83 ค่าเฉลี่ยความระลึก เท่ากับร้อยละ 52.49 และค่าเฉลี่ยความถ่วงดุล เท่ากับร้อยละ 62.04 และ Stacking มีค่าเฉลี่ยความแม่น เท่ากับร้อยละ 99.51 ค่าเฉลี่ยความเที่ยง เท่ากับร้อยละ 81.20 ค่าเฉลี่ยความระลึก เท่ากับร้อยละ 78.31 และค่าเฉลี่ยความถ่วงดุล ร้อยละ 79.73

แสดงผลการประเมินและเปรียบเทียบความแม่นของแต่ละอัลกอริทึมแบบรวมกลุ่ม ระหว่างเทคนิค Voting และเทคนิค Stacking ดังรูปที่ 13 และรูปที่ 14 ตามลำดับ



รูปที่ 13 : ผลการประเมินและเปรียบเทียบความแม่นของแต่ละอัลกอริทึมแบบรวมกลุ่มด้วยเทคนิค Voting



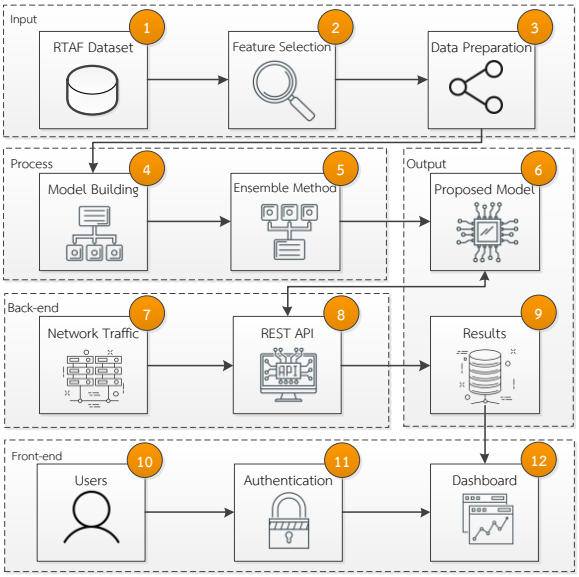
รูปที่ 14 : ผลการประเมินและเปรียบเทียบความแม่นของแต่ละอัลกอริทึมแบบรวมกลุ่มด้วยเทคนิค Stacking

จากผลการประเมินและเปรียบเทียบความแม่นของแต่ละอัลกอริทึมแบบรวมกลุ่มด้วยเทคนิค Stacking พบว่าอัลกอริทึม Gradient Boosted Trees ทำงานร่วมกับอัลกอริทึม Naïve Bayes ให้ประสิทธิภาพสูงสุด มีค่าเฉลี่ยความแม่น ร้อยละ 99.77 และค่าเฉลี่ยความเที่ยง ร้อยละ 88.59 ส่วนเทคนิค Voting ก็ให้ประสิทธิภาพสูงสุดเช่นกันเมื่อใช้อัลกอริทึม Gradient Boosted Trees ทำงานร่วมกับอัลกอริทึม Naïve Bayes มีค่าเฉลี่ยความแม่น ร้อยละ 98.99 และค่าเฉลี่ยความเที่ยง ร้อยละ 80.03 ดังนั้นการสร้างโมเดลแบบรวมกลุ่มด้วยอัลกอริทึม Gradient Boosted Trees ทำงานร่วมกับอัลกอริทึม Naïve Bayes ด้วยเทคนิค Stacking ให้ผลการประเมินประสิทธิภาพสูงสุดในการทำนายผลเมื่อเปรียบเทียบกับเทคนิคอื่น ๆ

จากนั้นผู้วิจัยจึงได้นำโมเดลที่ผ่านการประเมินและเปรียบเทียบประสิทธิภาพสูงสุด ไปทำสร้างเป็นระบบตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่องสำหรับเสริมสร้างความแข็งแกร่งและมีประสิทธิภาพด้านการรักษาความมั่นคงปลอดภัยไซเบอร์ในองค์กร ตามกระบวนการพัฒนาซอฟต์แวร์ (SDLC)

5.2) ผลการออกแบบระบบตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่อง

ผู้วิจัยนำโมเดลที่มีผลการประเมินและเปรียบเทียบประสิทธิภาพโมเดลสูงสุดในการวิจัยในครั้งนี้ พบว่าโมเดลตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่องแบบรวมกลุ่มระหว่างอัลกอริทึม Gradient Boosted Trees กับอัลกอริทึม Naïve Bayes ด้วยเทคนิค Stacking ให้ผลการประเมินและเปรียบเทียบประสิทธิภาพโมเดลสูงสุด ผู้วิจัยจึงนำโมเดลดังกล่าวไปสร้างเป็นระบบตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่อง สำหรับเสริมสร้างความแข็งแกร่งและมีประสิทธิภาพด้านการรักษาความมั่นคงปลอดภัยไซเบอร์ในองค์กร ซึ่งในการวิจัยในครั้งนี้ผู้วิจัยได้ออกแบบระบบ ดังรูปที่ 15



รูปที่ 15 : ผลการออกแบบระบบตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่อง

จากรูปที่ 15 แสดงความสัมพันธ์ของกระบวนการตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่องสำหรับเสริมสร้างความแข็งแกร่งและมีประสิทธิภาพด้านการรักษาความมั่นคงปลอดภัยไซเบอร์ในองค์กร อย่างไรก็ดี ระบบนี้ประกอบด้วยส่วนประกอบที่มีหน้าที่แตกต่างกัน 5 ส่วนดังต่อไปนี้

ส่วนที่ 1 Input คือ ขั้นตอนการนำข้อมูลที่ได้จากการดักจับข้อมูลการจราจรทางคอมพิวเตอร์จากระบบตรวจจับการบุกรุกทางไซเบอร์ (IDS) ของศูนย์ไซเบอร์กองทัพอากาศเข้าสู่ระบบ (หมายเลข 1) จากนั้นทำการคัดเลือกคุณลักษณะข้อมูล (feature selection) เพื่อทำการเลือกข้อมูลที่มีคุณลักษณะที่ตรงความต้องการของการออกแบบ (หมายเลข 2) จากนั้นทำการแบ่งกลุ่มข้อมูล (data preparation) เพื่อทำการทดสอบด้วยวิธี 10-fold cross-validation ซึ่งจะแบ่งข้อมูลออกเป็น 2 ส่วนคือ Train Set และ Test Set จำนวน 10 ส่วนเท่า ๆ กัน โดยการเปลี่ยนชุดข้อมูลทดสอบ ตั้งแต่ส่วนที่ 1 เป็นชุดทดสอบและส่วนที่ 2 ถึง 9 เป็นชุดสำหรับการเรียนรู้ไปจนถึงชุดทดสอบส่วนที่ 10 และส่วนที่ 1 ถึง 9 เป็นชุดสำหรับการเรียนรู้ เมื่อทำจนครบ 10 ครั้ง และนำค่าความแม่นของการ พยากาณ์ที่ได้ทั้งหมดมาหาค่าเฉลี่ย (หมายเลข 3)

ส่วนที่ 2 Process เป็นกระบวนการสร้างโมเดลตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่องด้วยโปรแกรม RapidMiner Studio จะใช้อัลกอริทึมการจำแนกข้อมูลการบุกรุก 4 อัลกอริทึม ประกอบด้วยอัลกอริทึม Decision Tree, Naïve Bayes, Random Forest และ Gradient Boosted Trees และประเมินและเปรียบเทียบความแม่นของแต่ละอัลกอริทึม (หมายเลข 4) จากนั้นสร้างโมเดลประเภทไฮบริด จากอัลกอริทึมที่มีผลประเมินและเปรียบเทียบความแม่นของแต่ละอัลกอริทึมที่สูงสุดไปไปสร้างเป็นโมเดลใหม่ด้วยเทคนิคการเรียนรู้ของเครื่องแบบรวมกลุ่ม และการประเมินและเปรียบเทียบประสิทธิภาพโมเดล (หมายเลข 5)

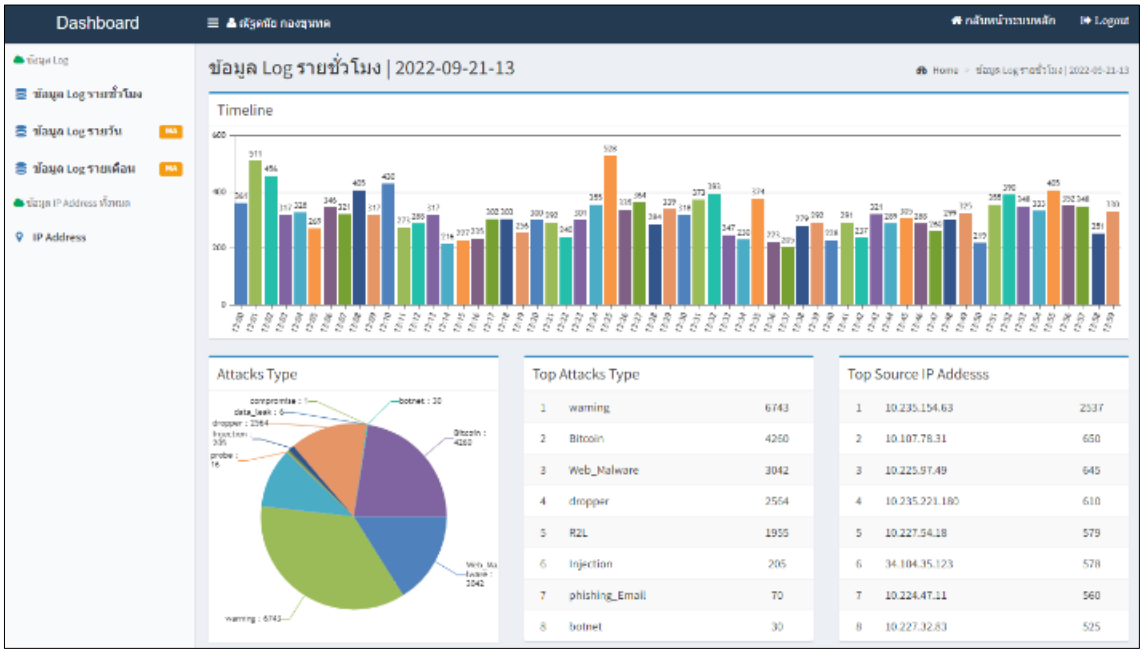
ส่วนที่ 3 Output เป็นกระบวนการนำโมเดลที่ทำการออกแบบและการประเมินและเปรียบเทียบประสิทธิภาพโมเดลเรียบร้อยแล้วให้อยู่ในรูปแบบพร้อมให้บริการ (หมายเลข 6) และสร้างฐานข้อมูลเพื่อเก็บผลการตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่องแบบรวมกลุ่ม (หมายเลข 9)

ส่วนที่ 4 Back-End เป็นกระบวนการดักจับข้อมูลการจราจรทางคอมพิวเตอร์จากระบบตรวจจับการบุกรุกทางไซเบอร์ (IDS) (หมายเลข 7) จากนั้นนำข้อมูลดังกล่าวส่งไปยังระบบแลกเปลี่ยนข้อมูล (REST API) (หมายเลข 8) เพื่อทำการวิเคราะห์ประเภทการบุกรุกทางไซเบอร์ เราจะพบว่า ระบบแลกเปลี่ยนข้อมูลจะทำหน้าที่เป็นตัวกลางในการแลกเปลี่ยนข้อมูลระหว่างโมเดลกับระบบแลกเปลี่ยนข้อมูลจากนั้นนำผลที่ได้จากการวิเคราะห์ประเภทการบุกรุกทางไซเบอร์ไปเก็บบันทึกไว้ในฐานข้อมูล (result) (หมายเลข 9) ซึ่งอยู่ในส่วนของ Output

ส่วนที่ 5 Front-End เป็นส่วนของผู้ใช้งานระบบ (หมายเลข 10) กล่าวคือ ผู้เข้าใช้งานระบบจะทำการยืนยันตัวตน (authentication) (หมายเลข 11) เพื่อตรวจสอบสิทธิ์การเข้าใช้งานระบบ จากนั้นถึงจะเข้าสู่หน้าหลัก (dashboard) (หมายเลข 12) ของระบบตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่องสำหรับเสริมสร้างความแข็งแกร่งและมีประสิทธิภาพด้านการรักษาความมั่นคงปลอดภัยไซเบอร์ ในองค์กร ดังรูปที่ 16

จากรูปที่ 16 แสดงหน้าระบบตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่อง ประกอบด้วยส่วนของข้อมูลการบุกรุกทางไซเบอร์ในรูปแบบเส้นเวลา (timeline) ซึ่งแสดงผลเป็นรายนาที ส่วนของประเภทการบุกรุกทางไซเบอร์ (attacks type) แสดงข้อมูลสถิติการวิเคราะห์การบุกรุกทางไซเบอร์โดยจำแนกตามประเภทการบุกรุกทางไซเบอร์

ส่วนแสดงภัยคุกตามทางไซเบอร์แยกตามจำนวนประเภทการบุกรุกทางไซเบอร์ (top attacks type) เรียงจากมากสุดไปน้อยสุดตามลำดับ และส่วนแสดงสถิติภัยคุกตามทางไซเบอร์แยกตามจำนวน IP Address ที่มีการบุกรุกทางไซเบอร์ (top source IP address) ที่เรียงจากมากสุดไปน้อยสุด ตามลำดับ



รูปที่ 16 : ระบบตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่อง

6) สรุปและอภิปรายผล

งานวิจัยนี้ผู้วิจัยได้ทำการเก็บรวบรวมข้อมูลจากระบบตรวจจับการบุกรุกทางไซเบอร์ ของศูนย์ไซเบอร์กองทัพอากาศ เพื่อจัดทำเป็นชุดข้อมูล จากนั้นทำการเตรียมข้อมูลให้เหมาะสมด้วยการเลือกข้อมูลที่มีความเหมาะสมไปจัดให้อยู่ในรูปแบบที่ถูกต้องเพื่อให้สามารถนำไปใช้ร่วมกับโปรแกรม RapidMiner Studio จากนั้นทำการแบ่งกลุ่มข้อมูลเพื่อทำการทดสอบด้วยวิธี 10-fold cross-validation ซึ่งจะแบ่งข้อมูลออกเป็น 2 ส่วนคือ Train Set และ Test Set จำนวน 10 ส่วนเท่าๆ กัน จากนั้นทำการสร้างโมเดลตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่องด้วยโปรแกรม RapidMiner Studio ด้วยอัลกอริทึมการจำแนกข้อมูลการบุกรุก 4 อัลกอริทึม ประกอบด้วยอัลกอริทึม Decision Tree, Naïve Bayes, Random Forest และ Gradient Boosted Trees เพื่อประเมินและเปรียบเทียบความแม่นของแต่ละอัลกอริทึมจะใช้สมการวัดค่าความแม่น และค่าความเที่ยงเป็นหลัก จากนั้นนำผลการประเมินและเปรียบเทียบประสิทธิภาพอัลกอริทึมที่ดีที่สุด คือ อัลกอริทึม Gradient Boosted Trees ไปสร้างเป็นอัลกอริทึมใหม่ด้วยเทคนิคการเรียนรู้ของเครื่องแบบรวมกลุ่ม กับอัลกอริทึม 3 อัลกอริทึม เพื่อประเมินและเปรียบเทียบความแม่นของแต่ละอัลกอริทึมแบบรวมกลุ่ม โดยใช้สมการวัดค่าความแม่น และค่าความเที่ยงที่ดีที่สุด

ผลการวิจัยพบว่า การสร้างโมเดลตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่องแบบรวมกลุ่มด้วยวิธีการนำอัลกอริทึม Gradient Boosted Trees ทำงานร่วมกับอัลกอริทึม Naïve Bayes ด้วยเทคนิค Stacking มีผลการประเมินและเปรียบเทียบประสิทธิภาพสูงสุดเมื่อเปรียบเทียบกับเทคนิคอื่นตามที่กล่าวมาข้างต้น จากนั้นนำอัลกอริทึม Gradient Boosted Trees กับอัลกอริทึม Naïve Bayes กับเทคนิค Stacking ไปสร้างเป็นระบบตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่อง สำหรับเสริมสร้างความแข็งแกร่งและมีประสิทธิภาพด้านการรักษาความมั่นคงปลอดภัยไซเบอร์ในองค์กร ประกอบด้วย 5 ส่วนหลัก ดังต่อไปนี้ 1) Input เป็นการนำข้อมูลที่ได้จากการดักจับข้อมูลการจราจรทางคอมพิวเตอร์จากระบบตรวจจับการบุกรุกทางไซเบอร์ของ ศูนย์ไซเบอร์กองทัพอากาศเข้าสู่ระบบ 2) Process เป็นการสร้างโมเดลตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่องด้วยโปรแกรม RapidMiner Studio 3) Output เป็นการนำโมเดลที่ทำการออกแบบและการประเมินและเปรียบเทียบประสิทธิภาพโมเดลเรียบร้อยแล้วให้อยู่ในรูปแบบพร้อมให้บริการ และสร้างฐานข้อมูลเพื่อเก็บผลการตรวจจับการบุกรุกทางไซเบอร์ 4) Back-End เป็นการดักจับข้อมูลการจราจรทางคอมพิวเตอร์จากระบบตรวจจับการบุกรุกทางไซเบอร์ จากนั้นนำข้อมูลส่งไปยังระบบแลกเปลี่ยนข้อมูล (REST API) เพื่อทำการวิเคราะห์ประเภทการบุกรุกทางไซเบอร์ จากนั้นระบบแลกเปลี่ยนข้อมูลจะทำหน้าที่เป็นตัวกลางในการแลกเปลี่ยนข้อมูลระหว่างโมเดลกับระบบแลกเปลี่ยนข้อมูลจากนั้นนำผลที่ได้จากการวิเคราะห์ประเภทการบุกรุกทางไซเบอร์ไปเก็บบันทึกไว้ในฐานข้อมูล ซึ่งอยู่ในส่วนของ Output และ 5) Front-End เป็นส่วนของผู้ใช้งานระบบนั่นคือ ผู้เข้าใช้งานระบบจะทำการยืนยันตัวตน (authentication) เพื่อตรวจสอบสิทธิ์การเข้าใช้งานระบบ จากนั้นถึงจะเข้าสู่หน้าหลัก (dashboard)

7) ข้อเสนอแนะ

จากที่ได้ทำการวิจัยเรื่องระบบตรวจจับการบุกรุกทางไซเบอร์แบบเรียลไทม์ด้วยการเรียนรู้ของเครื่อง สำหรับเสริมสร้างความแข็งแกร่งและมีประสิทธิภาพด้านการรักษาความมั่นคงปลอดภัยไซเบอร์ในองค์กร ผู้วิจัยได้มีข้อเสนอแนะเพิ่มเติมคือ ในการออกแบบโมเดลแบบรวมกลุ่ม ผู้วิจัยได้เลือกใช้เทคนิคการสร้างโมเดลแบบรวมกลุ่มด้วยเทคนิค Voting และ Stacking อย่างไรก็ตามยังคงมีอีกหลายเทคนิคที่นักวิจัยที่สนใจสามารถนำมาใช้ในการสร้างโมเดลแบบรวมกลุ่มได้อีกในอนาคต

กิตติกรรมประกาศ

ขอขอบพระคุณ ศูนย์ไซเบอร์กองทัพอากาศ ที่ได้สนับสนุนข้อมูลการจราจรทางคอมพิวเตอร์ จากระบบตรวจจับการบุกรุกทางไซเบอร์ เพื่อจัดทำเป็นชุดข้อมูลการตรวจจับการบุกรุกทางไซเบอร์ของกองทัพอากาศ และขอกราบขอบคุณ ท่านอาจารย์ที่ปรึกษา พลอากาศตรี ศาสตราจารย์ ดร.ประสงค์ ปราณีตพลกรัง สำนักบัณฑิตศึกษา โรงเรียนนายเรืออากาศนวมินทกษัตริยาธิราช และคณะทำงานที่ให้คำปรึกษา สนับสนุน และแนะนำตลอดระยะเวลาในการทำงานวิจัยครั้งนี้ จนสำเร็จได้อย่างสมบูรณ์

REFERENCES

[1] National Cyber Security Agency, “Cyber Threat Statistics,” 2023. [Online]. Available https://ncsa.or.th/service-statistics.html

[2] Royal Thai Air Force, "Royal Thai air force strategy for 20 years (B.E. 2561–2580) (revised version)," 2020. [Online]. Available: https://welcome-page.rtaf.mi.th/blog/e-ksaarephyaephr-11/yuththsaastrk-ngthaph-aakaas-20-pii-ph-s-2561-2580-38

[3] A. H. Janabi, T. Kanakis, and M. Johnson, "Overhead reduction technique for software-defined network based intrusion detection systems," *IEEE Access*, vol. 10, pp. 66481–66491, 2022.

[4] A. S. A. Aziz, S. E.-O. Hanafi, and A. E. Hassanien, "Comparison of classification techniques applied for network intrusion detection and classification," *J. Appl. Logic*, vol. 24, pp. 109–118, Nov. 2017.

[5] G. Karatas, O. Demir, and O. K. Sahingoz, "Increasing the performance of machine learning-based IDSs on an imbalanced and up-to-date dataset," *IEEE Access*, vol. 8, pp. 32150–32162, 2020.

[6] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: techniques, datasets and challenges," *Cybersecurity*, vol. 2, no. 1, 2019, Art. no. 20, doi: 10.1186/s42400-019-0038-7.

[7] T. A. Alamiedy, M. Anbar, A. K. Al-Ani, B. N. Al-Tamimi, and N. Faleh, "Review on feature selection algorithms for anomaly-based intrusion detection system," in *Proc. 3rd Int. Conf. Reliable Inf. and Commun. Technol.*, Kuala Lumpur, Malaysia, Jun. 2018, pp. 605–619.

[8] I. Garcia-Magarino, G. Gray, R. Lacuesta, and J. Lloret, "Survivability strategies for emerging wireless networks with data mining techniques: a case study with NetLogo and RapidMiner," *IEEE Access*, vol. 6, pp. 27958–27970, 2018.

[9] A. Karim, M. Shahroz, K. Mustofa, S. B. Belhaouari, and S. R. K. Joga, "Phishing detection system through hybrid machine learning based on URL," *IEEE Access*, vol. 11, pp. 36805–36822, 2023.

[10] T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.

[11] H. Yin, M. Xue, Y. Xiao, K. Xia, and G. Yu, "Intrusion detection classification model on an improved k-dependence Bayesian network," *IEEE Access*, vol. 7, pp. 157555–157563, 2019.

[12] L. Jian and K.-T. Chau, "Analytical calculation of magnetic field distribution in coaxial magnetic gears," *Prog. Electromagn. Res.*, vol. 92, pp. 1–16, 2009.

[13] N. Mahdi Abdulkareem and A. M. Abdulazeez, "Machine learning classification based on random forest algorithm: A review," *Int. J. Sci. Bus*., vol. 5, no. 2, pp. 128–142, 2021.

[14] W. Li, W. Wang, and W. Huo, "RegBoost: A gradient boosted multivariate regression algorithm," *Int. J. Crowd Sci.*, vol. 4, no. 1, pp. 60–72, 2020.

[15] D. Stiawan, "An approach for optimizing ensemble intrusion detection systems," *IEEE Access*, vol. 9, pp. 6930–6947, 2021.

[16] H. Luo, F. Cheng, H. Yu, and Y. Yi, "SDTR: Soft decision tree regressor for tabular data," *IEEE Access*, vol. 9, pp. 55999–56011, 2021.

[17] M. Mohy-Eddine, A. Guezzaz, S. Benkirane, M. Azrour, and Y. Farhaoui, "An ensemble learning based intrusion detection model for industrial IoT security," *Big Data Mining and Analytics*, vol. 6, no. 3, pp. 273–287, 2023.

[18] S. Ismail, Z. El Mrabet, and H. Reza, "An ensemble-based machine learning approach for cyber-attacks detection in wireless sensor networks," *Appl. Sci.*, vol. 13, no. 1, 2022, Art. no. 30.

[19] A. Agarwal, A. Agarwal, D. K. Verma, D. Tiwari, and R. Pandey, "A review on software development life cycle," *Int. J. Scientific Res. Comput. Sci., Eng. Inf. Technol.*, vol. 9, no. 3, pp. 384–388, 2023, doi: 10.32628/cseit2390387.

[20] C. Qi, J. Diao, and L. Qiu, "On estimating model in feature selection with cross-validation," *IEEE Access*, vol. 7, pp. 33454–33463, 2019.

[21] G.-P. Fernando, A.-A. H. Brayan, A. M. Florina, C.-B. Liliana, A.-M. Héctor-Gabriel, and T.-S. Reinel, "Enhancing intrusion detection in IoT communications through ML model generalization with a new dataset (IDSAI)," *IEEE Access*, vol. 11, pp. 70542–70559, 2023.

[22] L. Zou, X. Luo, Y. Zhang, X. Yang, and X. Wang, "HC-DTTSVM: A network intrusion detection method based on decision tree twin support vector machine and hierarchical clustering," *IEEE Access*, vol. 11, pp. 21404–21416, 2023.

[23] S. Ameer, "Comparative analysis of machine learning techniques for predicting air quality in smart cities," *IEEE Access*, vol. 7, pp. 128325–128338, 2019.

[24] W. Alhakami, A. ALharbi, S. Bourouis, R. Alroobaea, and N. Bouguila, "Network anomaly intrusion detection using a nonparametric Bayesian approach and feature selection," *IEEE Access*, vol. 7, pp. 52181–52190, 2019.

[25] X. Gao, C. Shan, C. Hu, Z. Niu, and Z. Liu, "An adaptive ensemble machine learning model for intrusion detection," *IEEE Access*, vol. 7, pp. 82512–82521, 2019.

[26] Y. Yu and N. Bian, "An intrusion detection method using few-shot learning," *IEEE Access*, vol. 8, pp. 49730–49740, 2020.

[27] Z. Wu, H. Zhang, P. Wang, and Z. Sun, "RTIDS: A robust transformer-based approach for intrusion detection system," *IEEE Access*, vol. 10, pp. 64375–64387, 2022.

[28] S. Gibson, B. Issac, L. Zhang, and S. M. Jacob, "Detecting spam email with machine learning optimized with bio-inspired metaheuristic algorithms," *IEEE Access*, vol. 8, pp. 187914–187932, 2020.

[29] R. Zhao, Y. Mu, L. Zou, and X. Wen, "A hybrid intrusion detection system based on feature selection and weighted stacking classifier," *IEEE Access*, vol. 10, pp. 71414–71426, 2022.

[30] A. A. Taha and S. J. Malebary, "An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine," *IEEE Access*, vol. 8, pp. 25579–25587, 2020.