

Received: 14 พ.ย. 2569

Revised: 20 ก.พ. 2569

Accepted: 23 ก.พ. 2569

การจำแนกประเภทยูอาร์แอลฟิชซิงโดยใช้การเรียนรู้ของเครื่อง  
ร่วมกับการปรับแต่งไฮเปอร์พารามิเตอร์ด้วยการค้นหาแบบกริด  
Machine Learning-Based Phishing URL Classification  
with Grid Search Hyperparameter Tuning

ตริยภรณ์ กองอ้น<sup>1</sup> และ นิติเศษฐ หมวดทองอ่อน<sup>1\*</sup>

<sup>1</sup>สาขาวิชาวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยสุโขทัยธรรมาธิราช นนทบุรี 11120 ประเทศไทย

Triyaporn Kongaon<sup>1</sup> and Nithizethe Mhuadthongon<sup>1\*</sup>

<sup>1</sup>School of Science and Technology, Sukhothai Thammathirat Open University,  
Nonthaburi 11120, Thailand

\*Corresponding author: Nithizethe.Mhu@stou.ac.th

## Abstract

The increasing use of the internet has led to new forms of data theft, which are considered a form of cybercrime. Phishing attacks are one method used to deceive users into disclosing personal information, and URL phishing is the most common type of such attacks. This research aims to present an approach for detecting phishing websites by developing a machine learning model for phishing URL classification, together with hyperparameter tuning using the grid search method across five algorithms, in order to obtain the model that provides the best phishing website classification performance. The evaluated algorithms include Neural Networks, Logistic Regression, Naive Bayes, Support Vector Machines using the SMO algorithm, and Decision Trees. The Weka software was used as the simulation and testing tool, and model performance was measured using five-fold cross validation. The results indicate that the Support Vector Machine model using the SMO algorithm achieved the best performance, with an accuracy of 95.55%, precision of 95.60%, recall of 95.60%, and an F1 score of 95.50%. These results demonstrate that the developed model can accurately classify phishing URLs and can serve as an important prototype for the effective future development of automated phishing detection systems.

**Keywords:** Machine Learning; URL Phishing; Support Vector Machine; Hyperparameter; Grid Search

## บทคัดย่อ

การใช้งานอินเทอร์เน็ตที่เพิ่มขึ้นอย่างต่อเนื่องก่อให้เกิดรูปแบบใหม่ของการขโมยข้อมูล ซึ่งถือเป็นรูปแบบของอาชญากรรมทางไซเบอร์ การโจมตีแบบฟิชชิ่งเป็นวิธีการหนึ่งที่มุ่งหลอกลวงผู้ใช้งานให้เปิดเผยข้อมูลส่วนบุคคลต่าง ๆ โดยการฟิชชิ่งผ่านยูอาร์แอลเป็นรูปแบบการโจมตีที่พบได้มากที่สุด งานวิจัยนี้มีวัตถุประสงค์เพื่อนำเสนอแนวทางในการตรวจจับเว็บไซต์ฟิชชิ่ง โดยพัฒนาแบบจำลองการเรียนรู้ของเครื่องสำหรับการจำแนกยูอาร์แอลฟิชชิ่งร่วมกับการปรับค่าไฮเปอร์พารามิเตอร์ด้วยวิธีการค้นหาแบบกริดผ่านอัลกอริทึม 5 วิธีการ เพื่อให้ได้มาซึ่งแบบจำลองที่ให้ผลการจำแนกเว็บไซต์ฟิชชิ่งที่ดีที่สุด ได้แก่ โครงข่ายประสาทเทียม การถดถอยโลจิสติก นาอ์ฟเบส ซัพพอร์ตเวกเตอร์แมชชีนที่ใช้อัลกอริทึม SMO และต้นไม้ตัดสินใจ โปรแกรม Weka ถูกใช้เป็นเครื่องมือในการจำลองและทดสอบ โดยทำการวัดประสิทธิภาพด้วยวิธีการตรวจสอบไขว้แบบ 5 ส่วน จากผลงานวิจัยพบว่า แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนที่ใช้อัลกอริทึม SMO ให้ผลลัพธ์ที่ดีที่สุด โดยมีค่าความถูกต้องเท่ากับ 95.55% ค่าความแม่นยำเท่ากับ 95.60% ค่าความระลึกเท่ากับ 95.60% และค่าความถ่วงดุลเท่ากับ 95.50% ผลงานวิจัยแสดงให้เห็นว่าแบบจำลองที่พัฒนาขึ้นสามารถจำแนกยูอาร์แอลฟิชชิ่งได้อย่างแม่นยำ สามารถใช้เป็นต้นแบบสำคัญสำหรับการพัฒนาระบบตรวจจับฟิชชิ่งแบบอัตโนมัติในอนาคตได้อย่างมีประสิทธิภาพ

**คำสำคัญ:** การเรียนรู้ของเครื่อง; ยูอาร์แอลฟิชชิ่ง; ซัพพอร์ตเวกเตอร์แมชชีน; ไฮเปอร์พารามิเตอร์; การค้นหาแบบกริด

## 1. บทนำ

ฟิชชิ่ง (phishing) เป็นการกระทำผิดทางไซเบอร์ที่ผู้ไม่ประสงค์ดีใช้กลวิธีหลอกลวงผู้ใช้งานให้เปิดเผยข้อมูลสำคัญ เช่น รหัสผ่านหรือข้อมูลทางการเงิน โดยหนึ่งในรูปแบบที่พบบ่อยที่สุดคือการสร้างเว็บไซต์ปลอมที่เลียนแบบเว็บไซต์จริง เช่น paypal หรือ eBay (Aljofey et al., 2020) การโจมตีในลักษณะดังกล่าวได้รับการจัดให้เป็นภัยคุกคามสำคัญในยุคดิจิทัล เนื่องจากมีแนวโน้มเพิ่มขึ้นอย่างต่อเนื่อง และส่งผลกระทบโดยตรงต่อความมั่นคงปลอดภัยของระบบสารสนเทศทั้งในภาคธุรกิจ และภาครัฐ ซึ่งสะท้อนให้เห็นถึงความจำเป็นในการพัฒนาแนวทางการตรวจจับที่มีประสิทธิภาพ และสามารถตอบสนองได้แบบอัตโนมัติ

เว็บไซต์ฟิชชิ่งในปัจจุบันมักใช้ยูอาร์แอล (uniform resource locator: URL) ที่มีลักษณะคล้ายกับเว็บไซต์จริงเพื่อหลอกให้ผู้ใช้งานเข้าสู่เว็บไซต์อันตรายที่มีเจตนาในการขโมยข้อมูลส่วนบุคคลหรือดาวนโหลดมัลแวร์ (Salloum et al., 2021; Ashok et al., 2024) อย่างไรก็ตามวิธีการตรวจจับแบบดั้งเดิม เช่น การใช้รายการเว็บไซต์ต้องห้าม (blacklist) หรือระบบตรวจสอบตามกฎเงื่อนไข (rule-based system) ยังไม่สามารถรับมือกับรูปแบบการโจมตีใหม่ ๆ ได้อย่างมีประสิทธิภาพ อีกทั้งยังมีอัตราการแจ้งเตือนผิดพลาดในระดับสูง ส่งผลให้ต้องอาศัยการตรวจสอบซ้ำจากผู้เชี่ยวชาญ และไม่สามารถนำไปใช้งานได้อย่างเหมาะสมในสภาพแวดล้อมจริง จึงเกิดช่องว่างของวิธีการตรวจจับที่ต้องอาศัยเทคนิคที่มีความยืดหยุ่น และสามารถเรียนรู้จากข้อมูลได้

จากข้อจำกัดดังกล่าว งานวิจัยนี้จึงเสนอแนวทางการตรวจจับเว็บไซต์ฟิชซิง โดยใช้การเรียนรู้ของเครื่อง (machine learning: ML) เนื่องจากสามารถเรียนรู้จากข้อมูลจำนวนมาก และจำแนกเว็บไซต์ฟิชซิงได้โดยอัตโนมัติ ซึ่งช่วยเพิ่มความแม่นยำ และลดความผิดพลาดจากการตรวจจับโดยมนุษย์ (Haq et al., 2024)

## 2. วัตถุประสงค์

1) เพื่อพัฒนาแบบจำลองการเรียนรู้ของเครื่องสำหรับการจำแนกยูอาร์แอลฟิชซิง โดยประยุกต์ใช้เทคนิคการปรับค่าไฮเปอร์พารามิเตอร์ด้วยวิธีการค้นหาแบบกริด เพื่อเพิ่มประสิทธิภาพ และความแม่นยำของแบบจำลอง

2) เพื่อเปรียบเทียบประสิทธิภาพของอัลกอริทึมการเรียนรู้ของเครื่องทั้ง 5 วิธี ได้แก่ โครงข่ายประสาทเทียม (NN) การถดถอยโลจิสติก (LR) นาอ์ฟเบส (NB) ซัพพอร์ตเวกเตอร์แมชชีน (SVM) ด้วยอัลกอริทึม SMO และต้นไม้ตัดสินใจ (DT) เพื่อระบุอัลกอริทึมที่ให้ประสิทธิภาพในการจำแนกเว็บไซต์ฟิชซิงได้ดีที่สุด

## 3. งานวิจัยที่เกี่ยวข้อง

งานวิจัยหลายฉบับได้ประยุกต์ใช้อัลกอริทึมการเรียนรู้ของเครื่องที่หลากหลาย เช่น Ahammad et al. (2022) ได้เสนอวิธีป่าสุ่ม (random forest: RF) ต้นไม้ตัดสินใจ (decision tree: DT) โลตัสจีบีเอ็ม (light gradient boosting machine: LightGBM) การถดถอยโลจิสติก (logistic regression: LR) และซัพพอร์ตเวกเตอร์แมชชีน (support vector machine: SVM) เพื่อใช้ในการสร้างแบบจำลองสำหรับการตรวจจับยูอาร์แอลที่เป็นอันตราย นอกจากนี้ Abutaha et al. (2021) ได้เสนอการวิเคราะห์เชิงภาษาของยูอาร์แอลร่วมกับเทคนิคการเรียนรู้ของเครื่อง และรายงานว่แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนให้ความแม่นยำมีค่าดีที่สุดเท่ากับ 99.89% Ademola and Boniface (2021) ได้ศึกษาการจำแนกคุณลักษณะของยูอาร์แอลฟิชซิงโดยใช้ SVM และนาอ์ฟเบส (Naïve Bayes: NB) โดยพบว่า SVM ยังมีข้อจำกัดด้านการปรับขนาดเมื่อใช้กับชุดข้อมูลขนาดใหญ่ ขณะที่ NB สามารถช่วยลดข้อจำกัดดังกล่าวได้ Rachid et al. (2020) ศึกษาพบว่า SVM มีประสิทธิภาพดีที่สุดในการจำแนกเว็บไซต์ฟิชซิงได้อย่างแม่นยำถึง 95.66% ขณะเดียวกัน Ghalechyan et al. (2024) ได้ประยุกต์ใช้ NN ในการจำแนกยูอาร์แอล และพบว่ายูอาร์แอลที่มีความยาวน้อยกว่า 50 ตัวอักษร จะมีค่าความแม่นยำเท่ากับ 97% ผลการศึกษาดังกล่าวสะท้อนให้เห็นว่าอัลกอริทึมการเรียนรู้ของเครื่องในกลุ่มต่าง ๆ มีศักยภาพสูงในการนำมาประยุกต์ใช้สำหรับการตรวจจับเว็บไซต์ฟิชซิง

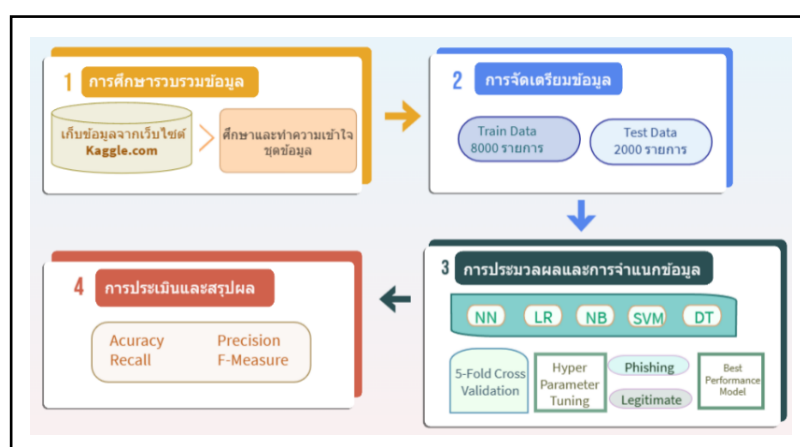
แม้ว่าอัลกอริทึมการเรียนรู้ของเครื่องจะมีศักยภาพสูง แต่ประสิทธิภาพของแบบจำลองยังคงขึ้นอยู่กับการกำหนดค่าไฮเปอร์พารามิเตอร์ที่เหมาะสม งานวิจัยที่เกี่ยวข้องชี้ให้เห็นว่างานจำนวนมากยังคงใช้ค่าไฮเปอร์พารามิเตอร์เริ่มต้นของโปรแกรมหรือมีการปรับค่าเพียงบางส่วนเท่านั้น Elgeldawi et al. (2021) ส่งผลให้แบบจำลองไม่สามารถสะท้อนศักยภาพสูงสุดของอัลกอริทึมได้ Anggreani et al. (2024) แสดงให้เห็นว่าการปรับค่าไฮเปอร์พารามิเตอร์ด้วยวิธีการค้นหาแบบกริดสามารถเพิ่มความแม่นยำของแบบจำลองได้อย่างมีนัยสำคัญ และ Justus et al. (2024) ได้เน้นย้ำว่าการปรับค่าไฮเปอร์พารามิเตอร์มีผลโดยตรงต่อประสิทธิภาพ

ของแบบจำลองในหลายมิติ จึงทำให้การปรับค่าไฮเปอร์พารามิเตอร์ด้วยวิธีการค้นหาแบบกริด เป็นแนวทางที่เหมาะสมสำหรับการเพิ่มประสิทธิภาพของแบบจำลอง

จากงานวิจัยที่เกี่ยวข้อง จึงมุ่งพัฒนา และเปรียบเทียบประสิทธิภาพของอัลกอริทึมการเรียนรู้ของเครื่องจำนวน 5 วิธี ได้แก่ โครงข่ายประสาทเทียม (NN) การถดถอยโลจิสติก (LR) นาอ์ฟเบส์ (NB) ซัพพอร์ตเวกเตอร์แมชชีน (SVM) ด้วยอัลกอริทึม SMO และต้นไม้ตัดสินใจ (DT) สำหรับการจำแนกยูอาร์แอลฟิชซิง โดยประยุกต์ใช้การปรับค่าไฮเปอร์พารามิเตอร์ด้วยวิธีการค้นหาแบบกริด เพื่อหาอัลกอริทึม และค่าพารามิเตอร์ที่ให้ประสิทธิภาพในการจำแนกยูอาร์แอลได้ดีที่สุด ซึ่งผลลัพธ์ที่ได้สามารถใช้เป็นแนวทางสำคัญในการพัฒนาเทคนิคการตรวจจับฟิชซิงที่มีความแม่นยำสูง และสามารถนำไปประยุกต์ใช้ได้จริงในระบบความปลอดภัยทางไซเบอร์ในอนาคต

#### 4. วิธีดำเนินงานวิจัย

งานวิจัยนี้ดำเนินการโดยเก็บรวบรวมข้อมูลเว็บไซต์ยูอาร์แอลที่เปิดเผยต่อสาธารณะจำนวน 10,000 รายการ เพื่อนำมาจัดเตรียมสำหรับการทดลอง โดยแบ่งข้อมูลออกเป็น 2 ชุด ได้แก่ ชุดข้อมูลสำหรับการฝึกสอน (training dataset) จำนวน 8,000 รายการ และชุดข้อมูลสำหรับการทดสอบ (testing dataset) จำนวน 2,000 รายการ ข้อมูลดังกล่าวถูกจัดหมวดหมู่ตามเงื่อนไขเป้าหมาย 2 กลุ่ม คือ กลุ่ม A เว็บไซต์ที่ถูกต้องตามกฎหมาย (legitimate) และกลุ่ม B เว็บไซต์ฟิชซิง (phishing) ทั้งนี้ ได้พัฒนาแบบจำลองด้วยวิธีการเรียนรู้ของเครื่อง 5 วิธี ได้แก่ โครงข่ายประสาทเทียม (NN) การถดถอยโลจิสติก (LR) นาอ์ฟเบส์ (NB) ซัพพอร์ตเวกเตอร์แมชชีน (SVM) ด้วยอัลกอริทึม SMO และต้นไม้ตัดสินใจ (DT) และใช้เทคนิคการตรวจสอบไขว้แบบ 5 ส่วน (5-fold cross validation) ในการประเมินความเสถียรของแบบจำลอง นอกจากนี้ ได้ปรับแต่งค่าไฮเปอร์พารามิเตอร์ของแต่ละอัลกอริทึมด้วยวิธีการค้นหาแบบกริด เพื่อระบุค่าที่เหมาะสมที่สุดของแต่ละแบบจำลอง ในขั้นตอนสุดท้าย ได้นำแบบจำลองที่ผ่านการฝึกสอนแล้วมาทดสอบกับชุดข้อมูลสำหรับการทดสอบอีกครั้ง เพื่อประเมินประสิทธิภาพของแต่ละแบบจำลอง โดยใช้ค่าจากคอนฟิวชันเมตริกซ์ (confusion matrix) ได้แก่ ค่าความถูกต้อง (accuracy) ค่าความแม่นยำ (precision) ค่าความระลึก (recall) และค่าความถ่วงดุล (F1-score) ตามลำดับ (นิสริวัฒน์ เจนศิริศักดิ์ และคณะ, 2025) กระบวนการดำเนินงานวิจัยแสดงไว้ในภาพที่ 1



ภาพที่ 1 วิธีดำเนินงานวิจัย

กระบวนการดำเนินงานวิจัยแบ่งออกได้เป็น 4 ขั้นตอน คือ

#### 4.1 การศึกษารวบรวมข้อมูล

ผู้วิจัยได้รวบรวมข้อมูลที่เกี่ยวข้องเพื่อใช้สำหรับการสร้างแบบจำลองการจำแนกยูอาร์แอลฟิชซิง เป็นข้อมูลที่ได้จาก web page phishing detection dataset ชุดข้อมูลนี้สร้างโดย (Hannousse et al., 2021) ถูกรวบรวมไว้ในเว็บไซต์ [www.kaggle.com](http://www.kaggle.com) และสามารถเข้าถึงได้อย่างเปิดเผยชื่อชุดข้อมูล “Dataset Phishing” จำนวน 11,430 รายการ ข้อมูลดังกล่าวถูกนำมาใช้เป็นชุดข้อมูลหลักในการสร้าง และประเมินแบบจำลองการเรียนรู้ของเครื่อง ทั้งนี้ เพื่อให้อัลกอริทึมสามารถประมวลผลได้อย่างมีประสิทธิภาพ จึงได้จัดเตรียมคุณลักษณะอินพุตทั้งหมด 87 ตัวแปร ซึ่งประกอบด้วยข้อมูลเชิงตัวเลข และข้อมูลเชิงข้อความ ที่ใช้ในการฝึกสอน และพัฒนาแบบจำลองการเรียนรู้ของเครื่อง ตัวอย่างโครงสร้างของชุดข้อมูลแสดงไว้ดังภาพที่ 2

| URL                                                          | length_url | length_host | ip | nb_dots | nb_hyphen | nb_at | nb_qm | nb_and |
|--------------------------------------------------------------|------------|-------------|----|---------|-----------|-------|-------|--------|
| http://174.139.46.123/ap/signin?openid.pape_max_auth_age=    | 477        | 14          | 1  | 24      | 0         | 1     | 1     | 9      |
| http://www.crestonwood.com/router.php                        | 37         | 19          | 0  | 3       | 0         | 0     | 0     | 0      |
| http://shadetretechnology.com/V4/validation/a111aedc8ae3     | 77         | 23          | 1  | 1       | 0         | 0     | 0     | 0      |
| https://support-apple.com.secureupdate.duilaawyork.com       | 126        | 50          | 1  | 4       | 1         | 0     | 1     | 2      |
| http://rgipt.ac.in                                           | 18         | 11          | 0  | 2       | 0         | 0     | 0     | 0      |
| http://www.iracing.com/track/gateway-motorsports-park/       | 55         | 15          | 0  | 2       | 2         | 0     | 0     | 0      |
| http://appleid.apple.com-app.es/                             | 32         | 24          | 0  | 3       | 1         | 0     | 0     | 0      |
| http://www.mutuo.it                                          | 19         | 12          | 0  | 2       | 0         | 0     | 0     | 0      |
| http://www.shadetretechnology.com/V4/validation/ba4b8bdd     | 81         | 27          | 1  | 2       | 0         | 0     | 0     | 0      |
| http://vamoastudiarmedicina.blogspot.com/                    | 42         | 34          | 0  | 2       | 0         | 0     | 0     | 0      |
| https://parade.com/425836/joshwagner/the-amazing-race-host   | 104        | 10          | 0  | 1       | 10        | 0     | 0     | 0      |
| https://www.astrologyonline.eu/Astro_MemoNew/Profilo.asp     | 56         | 22          | 0  | 3       | 0         | 0     | 0     | 0      |
| https://www.lifewire.com/tcp-port-21-818146                  | 43         | 16          | 0  | 2       | 3         | 0     | 0     | 0      |
| https://technofizi.net/top-best-mp3-downloader-app-for-andro | 83         | 14          | 0  | 1       | 9         | 0     | 0     | 0      |
| http://html.house/7cecid6.html                               | 31         | 10          | 0  | 2       | 0         | 0     | 0     | 0      |
| https://www.missfiga.com/                                    | 25         | 16          | 0  | 2       | 0         | 0     | 0     | 0      |
| http://wave.progressfilm.co.uk/time3/?logon=myposte          | 51         | 23          | 0  | 3       | 0         | 0     | 1     | 0      |
| https://www.chiefarchitect.com/                              | 31         | 22          | 0  | 2       | 0         | 0     | 0     | 0      |
| http://beta.kenaidanceta.com/postamok/d39a2/source           | 50         | 21          | 0  | 2       | 0         | 0     | 0     | 0      |
| http://www.ktplasmachinery.com/cs/                           | 34         | 23          | 0  | 2       | 0         | 0     | 0     | 0      |
| http://www.2345daohang.com/                                  | 27         | 19          | 0  | 2       | 0         | 0     | 0     | 0      |
| http://www.game.co.uk/en/games/nintendo-switch/nintendo-s    | 63         | 14          | 0  | 3       | 2         | 0     | 0     | 0      |
| https://blog.hubspot.com/marketing/email-open-click-rate-ber | 66         | 16          | 0  | 2       | 4         | 0     | 0     | 0      |
| http://batvrms.net/deliver/D2017HL/u.php                     | 40         | 11          | 0  | 2       | 0         | 0     | 0     | 0      |
| http://sophie-world.com/games/port-and-starboard             | 48         | 16          | 0  | 1       | 3         | 0     | 0     | 0      |
| https://support-apple.com.secureupdate.duilaawyork.com/      | 126        | 50          | 1  | 4       | 1         | 0     | 1     | 2      |
| https://polarklimatsgserver.blogspot.com/                    | 41         | 32          | 0  | 2       | 0         | 0     | 0     | 0      |
| https://www.provenancevineyards.com/                         | 36         | 27          | 0  | 2       | 0         | 0     | 0     | 0      |
| http://www.online-tech-tips.com/computer-tips/how-to-change  | 72         | 24          | 0  | 2       | 7         | 0     | 0     | 0      |

ภาพที่ 2 ตัวอย่างรายการชุดข้อมูลยูอาร์แอล จากชุดข้อมูล

web page phishing detection Hannousse, et al. (2021)

จากภาพที่ 2 แสดงให้เห็นตัวอย่างของรายการยูอาร์แอล และคุณลักษณะต่าง ๆ ที่ผ่านกระบวนการสกัด และวิเคราะห์เบื้องต้น เพื่อนำมาจัดกลุ่มข้อมูลออกเป็น 3 กลุ่มคุณลักษณะหลัก (feature classes) ได้แก่ กลุ่มที่ 1 คุณลักษณะที่ได้จากโครงสร้าง และไวยากรณ์ของยูอาร์แอล จำนวน 56 คุณลักษณะ กลุ่มที่ 2 คุณลักษณะที่ได้จากเนื้อหาภายในของยูอาร์แอล จำนวน 24 คุณลักษณะ และกลุ่มที่ 3 คุณลักษณะที่ได้จากการสืบค้นบริการภายนอก จำนวน 7 คุณลักษณะ โดยคุณลักษณะทั้ง 3 กลุ่มนี้ถูกนำมาใช้ร่วมกันในกระบวนการเรียนรู้ของเครื่อง เพื่อให้สามารถจำแนกเว็บไซต์ฟิชซิงออกจากเว็บไซต์ที่ถูกต้องตามกฎหมายได้อย่างมีประสิทธิภาพสูงสุด ทั้งนี้ คำอธิบายคุณลักษณะของชุดข้อมูล (data description) ทั้ง 3 กลุ่ม สามารถแสดงไว้ในภาพที่ 3 - 5 ตามลำดับ

| length_url                             | length_hostname                                         | ip                                                           | nb_dots                                                         | nb_hyphens                                                        | random_domain                                                       |
|----------------------------------------|---------------------------------------------------------|--------------------------------------------------------------|-----------------------------------------------------------------|-------------------------------------------------------------------|---------------------------------------------------------------------|
| ความยาวของ URL                         | ความยาวของชื่อโฮสต์                                     | มี IP อยู่ใน Hostname หรือไม่ (ส่งคืน 1 หรือ 0)              | จำนวนเครื่องหมาย (.) ใน URL                                     | จำนวนเครื่องหมาย (-) ใน URL                                       | มีการสุ่มเป็นตัวเลขด้วยตัวอักษรแบบ Random หรือไม่ (ส่งคืน 1 หรือ 0) |
| nb_and                                 | nb_or                                                   | nb_eq                                                        | nb_underscore                                                   | nb_tilde                                                          | nb_redirection                                                      |
| จำนวนเครื่องหมาย (&) ใน URL            | จำนวน เครื่องหมาย ( ) ใน URL                            | จำนวนเครื่องหมาย (=) ใน URL                                  | จำนวนเครื่องหมาย (.) ใน URL                                     | ตรวจสอบว่ามีสัญลักษณ์ (~) อยู่ใน URL หรือไม่ (ส่งคืน 1 หรือ 0)    | จำนวนของการเปลี่ยนเส้นทาง                                           |
| nb_star                                | nb_colon                                                | nb_comma                                                     | nb_semicolumn                                                   | nb_dollar                                                         | longest_words_raw                                                   |
| มีจำนวนเครื่องหมาย (*) ใน URL          | มีจำนวนเครื่องหมาย (:) ใน URL                           | มีจำนวนเครื่องหมาย (,) ใน URL                                | มีจำนวนเครื่องหมาย (;) ใน URL                                   | มีจำนวนเครื่องหมาย (\$) ใน URL                                    | ความยาวคำยาวที่สุด ในรายการ Raw Words                               |
| nb_com                                 | nb_slash                                                | http_in_path                                                 | https_token                                                     | ratio_digits_url                                                  | shortest_words_raw                                                  |
| มีจำนวน com ใน URL                     | มีจำนวนเครื่องหมาย (/) ใน URL                           | มีจำนวน http หรือ https                                      | มีการใช้ https หรือไม่ (ส่งคืน 1 หรือ 0)                        | มีจำนวนตัวเลขใน URL อยู่ในรูปแบบ โดเมนย่อย และ Path               | ความยาวคำสั้นที่สุด ในรายการ Raw Words                              |
| nb_at                                  | nb_gm                                                   | port                                                         | nb_subdomains                                                   | prefix_suffix                                                     | longest_word_path                                                   |
| จำนวนเครื่องหมาย (@) ใน URL            | จำนวนเครื่องหมาย (?) ใน URL                             | มี port อยู่ใน Domain หรือไม่ (ส่งคืน 1 หรือ 0)              | มีจำนวนโดเมนย่อย                                                | มี prefix /suffix หรือไม่ (ส่งคืน 1 หรือ 0)                       | ความยาวคำยาวที่สุด ใน Path                                          |
| nb_percent                             | nb_slash                                                | tid_in_path                                                  | shortening_service                                              | path_extension                                                    | nb_external_redirection                                             |
| มีจำนวนเครื่องหมาย (%) ใน URL          | มีจำนวนเครื่องหมาย (/) ใน URL                           | มีการใช้ TLD ใน Path หรือไม่ (ส่งคืน 1 หรือ 0)               | มีการใช้ URL หรือ IP (ส่งคืน 1 หรือ 0)                          | มี Path extension ที่เกี่ยวข้องด้วย tid หรือไม่ (ส่งคืน 1 หรือ 0) | จำนวนของการเปลี่ยนเส้นทางไปยังโดเมนอื่น                             |
| nb_space                               | nb_www                                                  | tid_in_subdomain                                             | shortest_word_host                                              | shortest_word_path                                                | domain_in_brand                                                     |
| มีจำนวนพื้นที่ว่าง (space, %20) ใน URL | มีจำนวน www ใน URL                                      | มีการใช้ TLD ใน โดเมนย่อยหรือไม่ (ส่งคืน 1 หรือ 0)           | ความยาวคำสั้นที่สุดในโฮสต์                                      | ความยาวคำสั้นที่สุดใน path                                        | มีโดเมนในรายการแบรนด์หรือไม่ (ส่งคืน 1 หรือ 0)                      |
| ratio_digits_host                      | puny_code                                               | abnormal_subdomain                                           | length_words_raw                                                | char_repeat                                                       | brand_in_path                                                       |
| อัตราส่วนของตัวเลขในชื่อโฮสต์          | ตรวจสอบว่ามี Puny Code ใน URL หรือไม่ (ส่งคืน 1 หรือ 0) | มีโดเมนย่อยที่ผิดปกติหรือไม่ เช่น www-wwNN (ส่งคืน 1 หรือ 0) | ความยาวของ Raw Words                                            | นับจำนวนตัวอักษรซ้ำ                                               | มีชื่อแบรนด์ใน path หรือไม่ (ส่งคืน 1 หรือ 0)                       |
|                                        | brand_in_subdomain                                      | avg_words_raw                                                | statistical_report                                              | phish_hints                                                       |                                                                     |
|                                        | มีชื่อแบรนด์ในโดเมนย่อยหรือไม่ (ส่งคืน 1 หรือ 0)        | นับความยาวคำเฉลี่ย ในรายการ Raw Words                        | รายงานทางสถิติ ว่ามีการรับ URL หรือ การรับ IP (ส่งคืน 1 หรือ 0) | จำนวนคำซ้ำใน path                                                 |                                                                     |
|                                        | suspicious_tid                                          | longest_word_host                                            | avg_word_host                                                   | avg_word_path                                                     |                                                                     |
|                                        | มี TLD ที่น่าสงสัย หรือไม่ (ส่งคืน 1 หรือ 0)            | ความยาวคำยาวที่สุดในโฮสต์                                    | ความยาวคำเฉลี่ยในโฮสต์                                          | ความยาวคำเฉลี่ยใน Path                                            |                                                                     |

ภาพที่ 3 คุณลักษณะของชุดข้อมูล กลุ่มที่ 1  
การสกัดจากโครงสร้าง และไวยากรณ์ของยูอาร์แอล จำนวน 56 คุณลักษณะ

| nb_hyperlinks                                 | ratio_intHyperlinks              | ratio_extHyperlinks                        | ratio_nullHyperlinks               | nb_extCSS                                              | right_click                                      |
|-----------------------------------------------|----------------------------------|--------------------------------------------|------------------------------------|--------------------------------------------------------|--------------------------------------------------|
| จำนวนไฮเปอร์ลิงก์ที่มีอยู่ในเว็บไซต์          | อัตราส่วนไฮเปอร์ลิงก์ภายใน       | อัตราส่วนไฮเปอร์ลิงก์ภายนอก                | อัตราส่วนไฮเปอร์ลิงก์ว่าง          | จำนวน CSS ภายนอก                                       | มีการกระทำ เมื่อคลิกขวาหรือไม่ (ส่งคืน 1 หรือ 0) |
| ratio_intRedirection                          | ratio_extRedirection             | ratio_intErrors                            | ratio_extErrors                    | login_form                                             | empty_title                                      |
| อัตราส่วนการเปลี่ยนเส้นทางภายใน               | อัตราส่วนการเปลี่ยนเส้นทางภายนอก | อัตราส่วนการสร้างข้อผิดพลาดภายใน           | อัตราส่วนการสร้างข้อผิดพลาดภายนอก  | มีฟอร์มล็อกอินหรือไม่ (ส่งคืน 1 หรือ 0)                | ตรวจสอบว่ามีชื่อว่างหรือไม่ (ส่งคืน 1 หรือ 0)    |
| external_favicon                              | links_in_tags                    | submit_email                               | ratio_intMedia                     | ratio_extMedia                                         | domain_in_title                                  |
| การมี favicon ภายนอกหรือไม่ (ส่งคืน 1 หรือ 0) | อัตราส่วนของจำนวนลิงก์ภายใน Tags | การส่งอีเมลไปอีจีมเมล                      | อัตราส่วนของสื่อภายใน              | อัตราส่วนของสื่อภายนอก                                 | มีโดเมนในหัวข้อหน้าหรือไม่ (ส่งคืน 1 หรือ 0)     |
| sfh                                           | iframe                           | popup_window                               | safe_anchor                        | onmouseover                                            | domain_with_copyright                            |
| ตัวจัดการแบบฟอร์มเซิร์ฟเวอร์                  | iframe สำหรับการเปลี่ยนเส้นทาง   | มีหน้าต่าง Popup หรือไม่ (ส่งคืน 1 หรือ 0) | อัตราส่วนของจุดเชื่อมโยงที่ปลอดภัย | มีการกระทำ เมื่อเมาส์ชี้ไปที่หรือไม่ (ส่งคืน 1 หรือ 0) | มีโดเมนที่ใส่ลิขสิทธิ์หรือไม่ (ส่งคืน 1 หรือ 0)  |

ภาพที่ 4 คุณลักษณะของชุดข้อมูล กลุ่มที่ 2  
คุณลักษณะที่ได้จากเนื้อหาภายในของยูอาร์แอล จำนวน 24 คุณลักษณะ

| whois_registered_domain                             | domain_registration_length                                  | domain_age                                  |
|-----------------------------------------------------|-------------------------------------------------------------|---------------------------------------------|
| โดเมนมีการรับรองโดย WHOIS หรือไม่ (ส่งคืน 1 หรือ 0) | อายุการจดทะเบียนโดเมน                                       | อายุโดเมนของ URL                            |
| web_traffic                                         | dns_record                                                  | google_index                                |
| จำนวนการเข้าใช้เว็บไซต์                             | มีระยะเวลาการหมดอายุของบันทึก DNS หรือไม่ (ส่งคืน 1 หรือ 0) | มีดัชนีของ Google หรือไม่ (ส่งคืน 1 หรือ 0) |
|                                                     | page_rank                                                   |                                             |
|                                                     | การจัดลำดับของเว็บไซต์                                      |                                             |

ภาพที่ 5 คุณลักษณะของชุดข้อมูล กลุ่มที่ 3  
คุณลักษณะที่ได้จากการสืบค้นบริการภายนอก จำนวน 7 คุณลักษณะ

## 4.2 การจัดเตรียมข้อมูล

ในขั้นตอนของการจัดเตรียมข้อมูลก่อนเข้าสู่กระบวนการวิเคราะห์ และการสร้างแบบจำลองการเรียนรู้ของเครื่องนั้น มีการพิจารณาแนวทางในการปรับลดจำนวนข้อมูล และลดความซับซ้อนของชุดข้อมูล เนื่องจากการลดมิติข้อมูล และการควบคุมขนาดของชุดข้อมูลถือเป็นเทคนิคพื้นฐานทางด้านการเรียนรู้ของเครื่องที่สามารถช่วยลดความซับซ้อนของข้อมูลอินพุต โดยจะเป็นการลดจำนวนตัวแปรหรือจำนวนข้อมูลที่ใช้ในการฝึกสอนแบบจำลอง ซึ่งมีบทบาทสำคัญต่อประสิทธิภาพทางด้านการคำนวณ และเสถียรภาพของแบบจำลอง (Stephen, 2024)

อย่างไรก็ตาม งานวิจัยที่เกี่ยวข้องได้ชี้ให้เห็นว่าการใช้ชุดข้อมูลที่มีขนาดใหญ่เกินความจำเป็น อาจนำไปสู่ภาวะโอเวอร์ฟิตติง (overfitting) และลดความสามารถในการสรุปผลของแบบจำลองได้ รวมถึงเพิ่มภาระด้านการคำนวณ ซึ่งอาจส่งผลให้การตรวจจับเว็บไซต์ฟิชซิงมีประสิทธิภาพที่ลดลง หรือไม่เหมาะสมต่อการนำไปประยุกต์ใช้งานจริง (Safi & Singh, 2023) ดังนั้น การกำหนดขนาดของชุดข้อมูลที่เหมาะสมจึงเป็นปัจจัยสำคัญในการสร้างแบบจำลองที่มีประสิทธิภาพ และจากงานวิจัยของ Ahammad et al. (2022) ได้ศึกษาการจำแนกยูอาร์แอลฟิชซิง โดยแบ่งข้อมูลข้อมูลสำหรับฝึกสอน 80% และข้อมูลสำหรับการทดสอบ 20% สามารถนำมาใช้ในการทำนายได้อย่างมีประสิทธิภาพ

เหตุผลดังกล่าว งานวิจัยนี้จึงได้ดำเนินการปรับลดจำนวนข้อมูลให้เหลือ 10,000 แถว โดยกำหนดให้ข้อมูลทั้งหมดมีป้ายกำกับผลลัพธ์ 2 ประเภท ได้แก่ legitimate (ยูอาร์แอลที่ถูกกฎหมาย) จำนวน 5,000 แถว และ phishing (ยูอาร์แอลฟิชซิง) จำนวน 5,000 แถว เพื่อใช้เป็นตัวแปรเป้าหมายสำหรับการจำแนกประเภทของแบบจำลอง และเพื่อให้ชุดข้อมูลมีความสมดุล (balanced dataset)

จากนั้น ผู้วิจัยได้แบ่งชุดข้อมูลออกเป็น 2 ส่วน ได้แก่ ชุดข้อมูลสำหรับการฝึกสอน (training dataset) จำนวน 8,000 แถว และชุดข้อมูลสำหรับการทดสอบ (testing dataset) จำนวน 2,000 แถว โดยข้อมูลทั้งสองชุดไม่มีการซ้ำกัน (non-overlapping datasets) เพื่อหลีกเลี่ยงความโน้มเอียงของแบบจำลอง และเพิ่มความน่าเชื่อถือของผลการทดลอง ข้อมูลทั้งหมดถูกจัดเก็บในรูปแบบไฟล์ CSV (comma-separated values) ซึ่งแต่ละคอลัมน์แทนคุณลักษณะของยูอาร์แอล และมีคอลัมน์ “status” สำหรับระบุผลลัพธ์ของการจำแนกประเภท โดยกำหนดให้ class a หมายถึง legitimate และ class b หมายถึง phishing จากนั้นได้ทำการแปลงชุดข้อมูลให้อยู่ในรูปแบบ ARFF (attribute relation file format) และถูกนำเข้าสู่อุปกรณ์ weka เพื่อใช้ในการทดลอง และการสร้างแบบจำลองการเรียนรู้ของเครื่องในขั้นตอนถัดไป

## 4.3 การประมวลผลและการจำแนกข้อมูล

การประมวลผล และการจำแนกข้อมูลในงานวิจัยนี้ ใช้กระบวนการเรียนรู้ของเครื่องเพื่อเปรียบเทียบประสิทธิภาพของอัลกอริทึมจำนวน 5 วิธี จำแนกข้อมูลได้แม่นยำที่สุด ในขั้นตอนนี้ใช้ข้อมูลจำนวน 10,000 แถว โดยแบ่งเป็นชุดข้อมูลสำหรับการฝึกสอน 80% หรือ 8,000 แถว และชุดข้อมูลสำหรับการทดสอบ 20% หรือ 2,000 แถว ข้อมูลทั้งหมดถูกนำเข้าสู่อุปกรณ์ weka เพื่อดำเนินการสร้างแบบจำลอง ทดสอบ และเปรียบเทียบผลลัพธ์ของแต่ละวิธี ระหว่างการฝึกสอนแบบจำลองได้ใช้เทคนิคการตรวจสอบไขว้แบบ 5 ส่วน เพื่อช่วยประเมินความเสถียรของแบบจำลอง และใช้วิธีการค้นหาแบบกริด เพื่อปรับค่าไฮเปอร์พารามิเตอร์ของแต่ละอัลกอริทึมให้มีความเหมาะสม

มากที่สุด หลังจากได้ค่าที่เหมาะสมมากที่สุดแล้ว จึงนำโมเดลดังกล่าว มาใช้สร้างแบบจำลองขั้นสุดท้าย และทำการทดสอบด้วยชุดข้อมูลสำหรับการทดสอบต่อไป สำหรับการประเมินประสิทธิภาพของแบบจำลองนั้น จะใช้ค่าที่คำนวณได้จาก confusion matrix ได้แก่ ค่าความถูกต้อง (accuracy) ค่าความแม่นยำ (precision) ค่าความระลึก (recall) และค่าความถ่วงดุล (F1-score) เพื่อเปรียบเทียบประสิทธิภาพของอัลกอริทึมทั้ง 5 วิธี และระบุแบบจำลองที่มีผลลัพธ์ที่ดีที่สุด

4.3.1 วิธีการค้นหาแบบกริด (grid search) เป็นวิธีแบบดั้งเดิมที่ใช้ในการเรียนรู้ของเครื่องสำหรับการปรับแต่งไฮเปอร์พารามิเตอร์ โดยจะทดลองค่าไฮเปอร์พารามิเตอร์ทุกชุดที่กำหนดไว้ เพื่อค้นหาโมเดลที่ดีที่สุด ซึ่งการค้นหาค่าพารามิเตอร์ที่เหมาะสมที่สุดของแบบจำลอง จะส่งผลให้ได้ผลการทำนายที่แม่นยำที่สุด

การสร้างแบบจำลองสำหรับการจำแนกข้อมูลในงานวิจัยนี้ จึงอาศัยผลการปรับค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมมากที่สุดของแต่ละอัลกอริทึม ซึ่งได้จากการค้นหาแบบกริด เพื่อให้แต่ละแบบจำลองมีประสิทธิภาพสูงสุดในการจำแนกยูอาร์แอลฟิชซึ่งออกจากยูอาร์แอลที่ถูกกฎหมาย โดยได้กำหนดค่าไฮเปอร์พารามิเตอร์เฉพาะของแต่ละอัลกอริทึม ได้แก่ วิธีนาอิวเบส (useKernelEstimator, useSuperviseDiscretization) วิธีซัพพอร์ตเวกเตอร์แมชชีน (C, kernel) วิธีต้นไม้ตัดสินใจ (confidenceFactor, minNumObj) วิธีโครงข่ายประสาทเทียม (hidden layer, learning rate) และวิธีการถดถอยโลจิสติก (maxIts)

4.3.2 วิธีนาอิวเบส (naïve bayes: NB) เป็นวิธีการจำแนกข้อมูลเชิงความน่าจะเป็นที่มีโครงสร้างเรียบง่าย และสามารถประมวลผลได้อย่างรวดเร็ว โดยมีสมมติฐานสำคัญว่าการเกิดของแต่ละคุณลักษณะเป็นอิสระต่อกัน ภายใต้สมมติฐานนี้ แบบจำลองจะทำการคำนวณความน่าจะเป็นของแต่ละคลาสเพื่อใช้ในการจำแนกข้อมูลใหม่ ทั้งนี้ สามารถอธิบายการคำนวณความน่าจะเป็นแบบมีเงื่อนไข ได้ดังสมการที่ (1)

$$p(h|d) = \frac{p(d|h) \times p(h)}{p(d)} \quad (1)$$

โดยที่  $p(h|d)$  เป็นความน่าจะเป็นที่จะเกิดเหตุการณ์  $h$  เมื่อเกิดเหตุการณ์  $d$

$p(d|h)$  เป็นความน่าจะเป็นที่จะเกิดเหตุการณ์  $d$  เมื่อเกิดเหตุการณ์  $h$

$p(h)$  เป็นความน่าจะเป็นที่จะเกิดเหตุการณ์  $h$

$p(d)$  เป็นความน่าจะเป็นที่จะเกิดเหตุการณ์  $d$

การสร้างโมเดลด้วยวิธี NB บนโปรแกรม weka ใช้อัลกอริทึม naive bayes และการกำหนดค่าไฮเปอร์พารามิเตอร์ด้วยการค้นหาแบบกริดของวิธี NB สามารถแสดงได้ดังตารางที่ 1

**ตารางที่ 1** ค่าไฮเปอร์พารามิเตอร์ของวิธี NB สำหรับการค้นหาแบบกริด

| ชื่อไฮเปอร์พารามิเตอร์     | ค่าไฮเปอร์พารามิเตอร์ |
|----------------------------|-----------------------|
| useKernelEstimator         | true, false           |
| useSuperviseDiscretization | true, false           |

โดยค่าไฮเปอร์พารามิเตอร์ที่นำมาพิจารณา ได้แก่ useKernelEstimator และ useSupervisedDiscretization โดย useKernelEstimator คือ พารามิเตอร์ที่ใช้ตัวประมาณค่าเคอร์เนล (Kernel Estimator) สำหรับการจัดการข้อมูลเชิงนาม โดยอาศัยอาร์กิวเมนต์ของเคอร์เนล

เพื่อประมาณการแจกแจงของข้อมูล ซึ่งช่วยเพิ่มความยืดหยุ่นในการประมาณความน่าจะเป็นของแอตทริบิวต์ ขณะที่ useSupervisedDiscretization คือ พารามิเตอร์ที่ใช้แปลงแอตทริบิวต์เชิงตัวเลขให้เป็นแอตทริบิวต์เชิงนามโดยอัตโนมัติภายใต้การเรียนรู้แบบมีผู้สอน เพื่อให้ข้อมูลมีความเหมาะสมต่อการสร้างแบบจำลองการจำแนกประเภทมากยิ่งขึ้น

4.3.3 วิธีซัพพอร์ตเวกเตอร์แมชชีน (support vector machine: SVM) เป็นอัลกอริทึมการเรียนรู้ของเครื่องที่สามารถประยุกต์ใช้ได้ทั้งการจำแนกข้อมูล และการถดถอย วิธีนี้มีหลักการสำคัญคือ การหาเส้นแบ่งข้อมูลที่ดีที่สุด (optimal hyperplane) โดยการเพิ่มระยะขอบ (margin) ระหว่างกลุ่มข้อมูลทั้งสองให้อยู่ห่างกันมากที่สุดในพื้นที่ของคุณลักษณะหรือพีเจอร์สเปซ ดังนั้น วิธี SVM จึงสามารถจำแนกข้อมูลที่มีหลายมิติได้อย่างมีประสิทธิภาพ สำหรับระยะขอบของวิธี SVM สามารถแสดงได้ดังสมการที่ (2) สำหรับค่าไฮเปอร์เพลนเชิงบวก และไฮเปอร์เพลนเชิงลบ สามารถแสดงได้ดังสมการที่ (3) และ (4) ตามลำดับ

$$w^T x + b = 0 \quad (2)$$

$$w^T x + b \geq 0 \text{ when } y_n = +1 \quad (3)$$

$$w^T x + b < 0 \text{ when } y_n = -1 \quad (4)$$

อัลกอริทึม SMO (sequential minimal optimization) เป็นอัลกอริทึมการปรับให้เหมาะสมซึ่งใช้ในการแก้ปัญหาที่เกิดขึ้นระหว่างการฝึกของ SVM ใช้แนวคิดการแก้ปัญหา quadratic programming: QP ผ่านการปรับตัวแปรเพียงสองตัวในแต่ละรอบ ส่งผลให้ลดเวลาและหน่วยความจำในการคำนวณ เหมาะสำหรับงานที่มีข้อมูลขนาดใหญ่ ซึ่ง SMO เป็นวิธีมาตรฐานที่ใช้ใน SVM ปัจจุบัน ซึ่งการทำงานของ SMO สามารถแสดงได้ดังสมการที่ (5)

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle \quad (5)$$

$$\text{subject to } 0 \leq \alpha_i \leq c, i = 1, \dots, m$$

$$\sum_{i=1}^m \alpha_i y^{(i)} = 0$$

การสร้างโมเดลด้วยวิธี SVM บนโปรแกรม weka ใช้อัลกอริทึม SMO และการกำหนดค่าไฮเปอร์พารามิเตอร์ด้วยการค้นหาแบบกริดของวิธี SVM สามารถแสดงได้ดังตารางที่ 2

ตารางที่ 2 ค่าไฮเปอร์พารามิเตอร์ของวิธี SVM สำหรับการค้นหาแบบกริด

| ชื่อไฮเปอร์พารามิเตอร์ | ค่าไฮเปอร์พารามิเตอร์                                                                               |
|------------------------|-----------------------------------------------------------------------------------------------------|
| C                      | [1, 5, 10, 15, 20, 25, 30]                                                                          |
| kernel                 | polykernel (exponent=1),<br>polykernel (exponent=2, 3),<br>rbfkernel (gamma = 0.01, 0.05, 0.1, 0.2) |

โดยค่าไฮเปอร์พารามิเตอร์ที่นำมาพิจารณา ได้แก่

C คือ พารามิเตอร์สำหรับควบคุมความซับซ้อนของแบบจำลอง โดยค่า C ที่สูงจะทำให้เส้นแบ่งการจำแนกมีระยะขอบแคบ และเพิ่มความซับซ้อนของแบบจำลอง ขณะที่ค่า C ที่ต่ำจะทำให้เส้นแบ่งมีระยะขอบกว้าง และช่วยลดความซับซ้อนของแบบจำลอง

kernel คือ ประเภทของฟังก์ชันทางคณิตศาสตร์ที่ใช้ในการแปลงข้อมูล โดยในงานวิจัยนี้พิจารณาเคอร์เนลเชิงเส้น (polykernel, exponent = 1) เคอร์เนลพหุนาม (polykernel, exponent = 2 และ 3) และเคอร์เนลฟังก์ชันฐานรัศมี (radial basis function kernel: RBFKernel)

exponent คือ พารามิเตอร์ของเคอร์เนลพหุนามที่ใช้กำหนดค่ายกกำลังของฟังก์ชัน ซึ่งมีผลต่อระดับความซับซ้อนและความโค้งของเส้นแบ่งการจำแนก

gamma คือ พารามิเตอร์ของเคอร์เนลฟังก์ชันฐานรัศมีที่ใช้กำหนดอิทธิพลของข้อมูลฝึกสอนต่อรูปร่างของเส้นแบ่งการจำแนก โดยค่า gamma ต่ำจะทำให้เส้นแบ่งมีความเรียบง่าย ขณะที่ค่า gamma สูงจะเพิ่มความจำเพาะของแบบจำลองต่อข้อมูลฝึกสอน

4.3.4 วิธีต้นไม้ตัดสินใจ (decision tree: DT) เป็นอัลกอริทึมที่สามารถประยุกต์ใช้ได้ทั้งการจำแนกข้อมูล และการถดถอย โดยมีหลักการทำงานเริ่มต้นจากโหนดราก (root node) ซึ่งนำค่าคุณลักษณะของข้อมูลไปเปรียบเทียบกับคุณลักษณะของโหนด เพื่อทำการตัดสินใจในการจำแนกกลุ่มข้อมูล การประมวลผลจะดำเนินไปตามลำดับชั้นของโหนดจนถึงโหนดใบ (leaf node) ซึ่งเป็นผลลัพธ์สุดท้ายของการจำแนกข้อมูล ต้นไม้ตัดสินใจที่ใช้สำหรับการจำแนก (classification tree) จะใช้ ค่า gini impurity หรือค่าเอนโทรปี (entropy) เป็นฟังก์ชันวัตถุประสงค์ (objective function) เพื่อหาจุดแบ่งข้อมูลที่ดีที่สุด (split point) ในแต่ละระดับของต้นไม้ (Chaiyaphop Jamjumrat, 2022)

การคำนวณ gini impurity คือ ค่าความน่าจะเป็น (likelihood) ของเหตุการณ์ที่สนใจ โดยเป็นค่าที่โมเดล decision tree ใช้ในการตัดสินใจเลือกโหนด และสร้างทางแยก (split) (Pasith Thanapatpisarn, 2022) ซึ่งสามารถแสดงได้ดังสมการที่ (6)

$$G = \sum_{k=1}^K \hat{p}_{mk} (1 - \hat{p}_{mk}) \quad (6)$$

โดยที่  $\hat{p}_{mk}$  คือ สัดส่วนของคลาส (class) ที่  $k$  ในโหนดที่  $m$

$K$  คือจำนวนคลาสทั้งหมด

$\sum_{k=1}^K$  คือ ผลรวมของค่าสำหรับทุกคลาสตั้งแต่  $k = 1$  ถึง  $K$

การสร้างโมเดลด้วยวิธี DT บนโปรแกรม weka ใช้อัลกอริทึม J48 และการกำหนดค่าไฮเปอร์พารามิเตอร์ด้วยการค้นหาแบบกริดของวิธี DT สามารถแสดงได้ดังตารางที่ 3

**ตารางที่ 3** ค่าไฮเปอร์พารามิเตอร์ของวิธี DT สำหรับการค้นหาแบบกริด

| ชื่อไฮเปอร์พารามิเตอร์ | ค่าไฮเปอร์พารามิเตอร์ |
|------------------------|-----------------------|
| confidenceFactor       | 0.25, 0.5, 0.75       |
| minNumObj              | 1, 2, 3, 4, 5         |

โดยค่าไฮเปอร์พารามิเตอร์ที่นำมาพิจารณา ได้แก่

confidenceFactor คือ พารามิเตอร์ที่ใช้กำหนดระดับช่วงความเชื่อมั่น (confidence interval) สำหรับการตัดแต่งกิ่งของต้นไม้ตัดสินใจ โดยค่า confidenceFactor ที่ต่ำจะทำให้เกิดการตัดแต่งกิ่งมากขึ้น ส่งผลให้แบบจำลองมีความเรียบง่าย ขณะที่ค่า confidenceFactor ที่สูงจะทำให้ต้นไม้มีความซับซ้อนมากขึ้น

minNumObj คือ จำนวนข้อมูลเรียนรู้ขั้นต่ำที่กำหนดให้มีอยู่ในโหนดใบ (leaf node) โดยหากการแบ่งโหนดทำให้จำนวนข้อมูลในโหนดใบต่ำกว่าค่าที่กำหนด โหนดดังกล่าวจะไม่ถูกนำไปใช้ในการตัดสินใจ ซึ่งช่วยลดความซับซ้อนของต้นไม้ และลดความเสี่ยงของการเกิดภาวะโอเวอร์ฟิตติง

4.3.5 วิธีโครงข่ายประสาทเทียม (neural network: NN) เป็นแบบจำลองทางคณิตศาสตร์ที่เลียนแบบการทำงานของระบบประสาทในสมองของมนุษย์ โดยจำลองกระบวนการส่งสัญญาณระหว่างเซลล์ประสาทเพื่อใช้ในการเรียนรู้ และตัดสินใจ แบบจำลองนี้ประกอบด้วย หน่วยประมวลผลย่อยที่เรียกว่า เซลล์ประสาทเทียม (artificial neuron) ซึ่งแต่ละเซลล์รับข้อมูลนำเข้าหลายตัวแปร จากนั้นทำการคำนวณน้ำหนัก และรวมค่าเพื่อสร้างเอาต์พุตออกมา โดยทั่วไป เซลล์ประสาทเทียมจะมีข้อมูลป้อนเข้าหลายตัวแปร ทำให้สามารถแสดงความสัมพันธ์ระหว่างอินพุต และเอาต์พุตได้ในรูปของสมการทางคณิตศาสตร์ เซลล์ประสาทเทียมที่ถูกจำลองขึ้นจะเป็นการนำข้อมูลป้อนเข้า ( $p$ ) คูณกับค่าถ่วงน้ำหนัก ( $w$ ) แล้วนำมารวมกับผลคูณของไบแอส ( $b$ ) กับตัวแปรต้นของไบแอส ซึ่งมีค่าเท่ากับ 1 โดยที่  $n$  คือ ตัวแปรต้นสุทธิ (net input) ดังสมการที่ (7)

$$n = wp + b \quad (7)$$

$a$  คือค่าตัวแปรตามของเซลล์ประสาท (neural output) โดยที่  $f$  คือ ฟังก์ชันการเปลี่ยนแปลง ดังสมการที่ (8)

$$a = f(wp + b) \quad (8)$$

โดยทั่วไปเซลล์ประสาทเทียม มีข้อมูลป้อนเข้าหรือตัวแปรต้นมากกว่าหนึ่งตัวแปร จึงทำให้ได้สมการเซลล์ประสาทเทียมดังสมการที่ (9) และ (10)

$$n = \sum (wp + b) \quad (9)$$

$$a = f(\sum wp + b) \quad (10)$$

การสร้างโมเดลด้วยวิธี NN บนโปรแกรม weka ใช้อัลกอริทึม MLP และการกำหนดค่าไฮเปอร์พารามิเตอร์ด้วยการค้นหาแบบกริดของวิธี NN สามารถแสดงได้ดังตารางที่ 4

**ตารางที่ 4** ค่าไฮเปอร์พารามิเตอร์ของวิธี NN สำหรับการค้นหาแบบกริด

| ชื่อไฮเปอร์พารามิเตอร์ | ค่าไฮเปอร์พารามิเตอร์       |
|------------------------|-----------------------------|
| hidden layer           | [1, 2, 3, 4, 5]             |
| learning rate          | [0.01, 0.05, 0.1, 0.2, 0.3] |

โดยไฮเปอร์พารามิเตอร์ที่นำมาพิจารณา ได้แก่

hidden layer คือ พารามิเตอร์ที่ใช้กำหนดจำนวนชั้นซ่อนของโครงข่ายประสาทเทียม โดยกำหนดค่าเริ่มต้นเป็น a (Auto) ซึ่งหมายถึงการให้ระบบปรับจำนวนโหนดอัตโนมัติ หรือกำหนดค่าเป็น 1 ชั้นซ่อน เพื่อใช้ในการเปรียบเทียบประสิทธิภาพของแบบจำลอง

learning rate คือ อัตราการเรียนรู้ที่ใช้ควบคุมขนาดของการปรับค่าน้ำหนักในแต่ละรอบการฝึกสอน โดยกำหนดจำนวนรอบการฝึกสอน (epoch) เท่ากับ 500 ซึ่งเป็นค่าเริ่มต้นของระบบ หมายถึงการฝึกสอนแบบจำลองซ้ำจำนวน 500 รอบ

4.3.6 วิธีการถดถอยโลจิสติก (logistic regression: LR) เป็นวิธีการทางสถิติที่ใช้สำหรับการพยากรณ์ความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ เมื่อเปรียบเทียบกับความน่าจะเป็นของการไม่เกิดเหตุการณ์ดังกล่าว ซึ่งการถดถอยแบบโลจิสติกจะมีตัวแปรตามเป็นตัวแปรเชิงคุณภาพ และสามารถแบ่งออกเป็นสองประเภทหลัก ได้แก่ การถดถอยโลจิสติกทวิภาค (binary logistic regression) ซึ่งตัวแปรตามมีค่าเพียงสองค่า และการถดถอยโลจิสติกพหุ (multinomial logistic regression) ซึ่งตัวแปรตามมีค่ามากกว่าสองค่าขึ้นไป

ในส่วนของตัวแปรอิสระ สามารถเป็นได้ทั้งตัวแปรเชิงปริมาณ ตัวแปรเชิงคุณภาพ หรือเป็นการผสมผสานของตัวแปรทั้งสองประเภท ฟังก์ชันโลจิสติก ถูกนำมาใช้เพื่ออธิบายความสัมพันธ์ระหว่างตัวแปรอิสระกับความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ โดยค่าความน่าจะเป็นจะถูกจำกัดให้อยู่ในช่วงระหว่าง 0 ถึง 1 ซึ่งสามารถแสดงได้ดังสมการที่ (11) และ (12) ตามลำดับ (รัชนิวรรณ ไพศาลวรรณเกียรติ, 2564)

$$prob(event) = \frac{e^{(\beta_0 + \beta_1 X)}}{1 + e^{(\beta_0 + \beta_1 X)}} \quad (11)$$

หรือ

$$prob(event) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} \quad (12)$$

เมื่อ  $\beta_0$  คือ ค่าคงที่ (เมื่อไม่มีอิทธิพลจากตัวแปรอิสระใด)

$\beta_1$  คือ ค่าสัมประสิทธิ์ของตัวแปรอิสระ (ประมาณได้จากข้อมูลสังเกต)

$X$  คือ ตัวแปรอิสระ (ตัวแปรทำนาย)

$e$  คือ ฟังก์ชันเอ็กซ์โพเนนเชียล (มีค่าประมาณ 2.71828)

การสร้างโมเดลด้วยวิธี LR บนโปรแกรม weka ใช้อัลกอริทึม logistic และการกำหนดค่าไฮเปอร์พารามิเตอร์ด้วยการค้นหาแบบกริดของวิธี LR สามารถแสดงได้ตามตารางที่ 5

ตารางที่ 5 ค่าไฮเปอร์พารามิเตอร์ของวิธี LR สำหรับการค้นหาแบบกริด

| ชื่อไฮเปอร์พารามิเตอร์ | ค่าไฮเปอร์พารามิเตอร์      |
|------------------------|----------------------------|
| maxIts                 | [-2, -1, 0, 1, 2, 3, 4, 5] |

โดยค่าไฮเปอร์พารามิเตอร์ที่นำมาพิจารณา ได้แก่

maxIts คือ พารามิเตอร์ที่ใช้กำหนด จำนวนรอบการวนซ้ำสูงสุด (maximum iterations) ในกระบวนการฝึกสอนแบบจำลอง โดยเมื่อถึงจำนวนรอบที่กำหนด กระบวนการฝึกสอนจะสิ้นสุดลง แม้ว่าแบบจำลองจะยังไม่สมบูรณ์

#### 4.4 การประเมินผล

การประเมินผลในงานวิจัยนี้ มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องทั้ง 5 วิธีที่ได้พัฒนาไว้ โดยทำการปรับแต่งค่าไฮเปอร์พารามิเตอร์ (hyperparameter tuning) ของแต่ละแบบจำลองให้มีความเหมาะสมมากที่สุด จากนั้น นำแบบจำลองที่ได้มาทดสอบกับชุดข้อมูลสำหรับการทดสอบ เพื่อประเมินความสามารถในการจำแนกข้อมูลได้อย่างถูกต้อง สำหรับการวิเคราะห์ผลการจำแนกนั้นจะใช้ confusion matrix ในการคำนวณเพื่อเปรียบเทียบผลลัพธ์การทำนายกับค่าจริง โดยพิจารณาจากค่าตัวชี้วัดประสิทธิภาพ 4 ค่า ได้แก่ ค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึกลับ และค่าความถ่วงดุล ซึ่งสามารถแสดงได้ดังสมการที่ 13 – 16 ตามลำดับ

ค่า accuracy คือ ค่าความถูกต้อง ที่ตัวแบบจำลองสามารถพยากรณ์ข้อมูลยูอาร์แอลฟิชชิ่งและยูอาร์แอลที่ถูกกฎหมาย ได้อย่างถูกต้องต่อข้อมูลทั้งหมด

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (13)$$

ค่า precision คือ ค่าความแม่นยำ ในการทำนายของกลุ่มเป้าหมาย

$$Precision = \frac{TP}{TP + FP} \quad (14)$$

ค่า recall คือ ค่าความระลึกลับ เป็นอัตราส่วนหรือจำนวนความถูกต้องในการจำแนกข้อมูลกลุ่มเป้าหมายเป็นจริง (true positive rate: TPR)

$$Recall = \frac{TP}{TP + FN} \quad (15)$$

ค่า F1-score คือ ค่าความถ่วงดุล การหาค่าเฉลี่ยของ precision และ recall

$$F1-Score = 2 \times \frac{precision \times recall}{precision + recall} \quad (16)$$

#### 5. ผลการทดลอง

งานวิจัยนี้มุ่งเน้นการพัฒนา และประเมินประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องสำหรับการจำแนกยูอาร์แอลฟิชชิ่ง ทั้งนี้ได้ดำเนินการประมวลผลข้อมูลภายใต้เงื่อนไขเดียวกัน เพื่อให้สามารถเปรียบเทียบประสิทธิภาพของแบบจำลองได้อย่างเหมาะสม โดยรายละเอียดของผลลัพธ์ที่ได้จากการทดลองการจำแนกยูอาร์แอลด้วยวิธีการเรียนรู้ของเครื่องทั้ง 5 วิธี แสดงดังต่อไปนี้

5.1 ผลการทดลองการจำแนกประเภทข้อมูลด้วยวิธี NN ด้วยอัลกอริทึม multilayer perceptron (MLP) ซึ่งเป็นโครงข่ายประสาทเทียมแบบหลายชั้นที่สามารถเรียนรู้ความสัมพันธ์ที่ซับซ้อนของข้อมูลได้ โดยเมื่อพิจารณาค่าความถูกต้อง จากการปรับค่าไฮเปอร์พารามิเตอร์ พบว่าค่าที่เหมาะสมซึ่งให้ค่าความถูกต้องสูงสุด ได้แก่ จำนวนชั้นซ่อน (hidden layer) เท่ากับ 1 และอัตราการเรียนรู้ (learning rate) เท่ากับ 0.01 ผลลัพธ์ของการทดลองสามารถแสดงได้ดังภาพที่ 6

| Hyperparameter |      | Training Time (epoch) = 500 |       |       |       |       |
|----------------|------|-----------------------------|-------|-------|-------|-------|
| Learning Rate  | 0.01 | 95.33                       | 94.21 | 94.18 | 94.31 | 94.55 |
|                | 0.05 | 94.78                       | 94.03 | 93.66 | 93.97 | 93.97 |
|                | 0.1  | 94.73                       | 93.95 | 93.89 | 94.02 | 94.17 |
|                | 0.2  | 94.41                       | 93.54 | 93.35 | 93.84 | 93.95 |
|                | 0.3  | 93.99                       | 93.8  | 93.44 | 94.16 | 93.86 |
| Hyperparameter |      | 1                           | 2     | 3     | 4     | 5     |
|                |      | Hidden Layer                |       |       |       |       |

ภาพที่ 6 การเปรียบเทียบการปรับแต่งค่าไฮเปอร์พารามิเตอร์ ด้วยอัลกอริทึม MLP

5.2 ผลการทดลองการจำแนกประเภทข้อมูลด้วย LR ด้วยอัลกอริทึม logistic ซึ่งเป็นอัลกอริทึมการจำแนกประเภทที่เหมาะสมสำหรับกรณีนี้ตัวแปรตามเป็นตัวแปรเชิงหมวดหมู่ โดยเมื่อพิจารณา ค่าความถูกต้อง จากการปรับค่าไฮเปอร์พารามิเตอร์ พบว่าค่าที่เหมาะสมซึ่งให้ค่าความถูกต้องสูงสุด ได้แก่ ค่า maxlts เท่ากับ -1 ซึ่งหมายถึงการกำหนดให้กระบวนการฝึกสอนทำงานจนกว่าจะเกิดการลู่เข้าของแบบจำลอง ผลลัพธ์ของการทดลองสามารถแสดงได้ดังภาพที่ 7

| Hyperparameter | maxlts |       |       |       |       |       |       |       |
|----------------|--------|-------|-------|-------|-------|-------|-------|-------|
|                | -2     | -1    | 0     | 1     | 2     | 3     | 4     | 5     |
| Accuracy       | 50.00  | 94.59 | 50.00 | 50.00 | 88.39 | 92.33 | 93.81 | 94.15 |

ภาพที่ 7 การเปรียบเทียบการปรับแต่งค่าไฮเปอร์พารามิเตอร์ ด้วยอัลกอริทึม logistic

5.3 ผลการทดลองการจำแนกประเภทข้อมูลด้วยวิธี NB ด้วยอัลกอริทึม naïve bayes ซึ่งเป็นอัลกอริทึมการจำแนกประเภทที่โดยทั่วไปสมมติให้ตัวแปรอินพุตเป็นตัวแปรเชิงนาม อย่างไรก็ตาม อัลกอริทึมดังกล่าว สามารถรองรับตัวแปรอินพุตเชิงตัวเลขได้โดยอาศัยสมมติฐานเกี่ยวกับการกระจายตัวของข้อมูล โดยเมื่อพิจารณา ค่าความถูกต้อง จากการปรับค่าไฮเปอร์พารามิเตอร์ พบว่าค่าที่เหมาะสมซึ่งให้ค่าความถูกต้องสูงสุด ได้แก่ useKernelEstimator เท่ากับ “False” และ useSupervisedDiscretization เท่ากับ “True” ผลลัพธ์ของการทดลองสามารถแสดงได้ดังภาพที่ 8

| Hyperparameter                                                    | Accuracy |
|-------------------------------------------------------------------|----------|
| useKernelEstimator = FALSE<br>useSupervisedDiscretization = FALSE | 81.05    |
| useKernelEstimator = TRUE<br>useSupervisedDiscretization = FALSE  | 89.41    |
| useKernelEstimator = FALSE<br>useSupervisedDiscretization = TRUE  | 91.14    |

ภาพที่ 8 การเปรียบเทียบการปรับแต่งค่าไฮเปอร์พารามิเตอร์ด้วยอัลกอริทึม Naïve Bayes

5.4 ผลการทดลองการจำแนกประเภทข้อมูลด้วยวิธี SVM ด้วยอัลกอริทึม SMO ซึ่งเป็นอัลกอริทึมที่ใช้ในการแก้ปัญหาการหาค่าเหมาะสมของ SVM โดยอาศัยไฮเปอร์พารามิเตอร์ต่าง ๆ เพื่อกำหนดลักษณะของเส้นแบ่งการจำแนก (hyperplane) โดยเมื่อพิจารณา ค่าความถูกต้อง จากการปรับค่าไฮเปอร์พารามิเตอร์ พบว่าค่าที่เหมาะสมซึ่งให้ค่าความถูกต้องสูงสุด ได้แก่ เคอร์เนลฟังก์ชันฐานรัศมี (RBFKernel) ที่มีค่า gamma เท่ากับ 0.1 และค่าพารามิเตอร์ C เท่ากับ 15 ผลลัพธ์ของการทดลองสามารถแสดงได้ดังภาพที่ 9

| Hyperparameter |    | Kernel     |       |       |           |       |       |       |
|----------------|----|------------|-------|-------|-----------|-------|-------|-------|
|                |    | PolyKernel |       |       | RBFKernel |       |       |       |
| c              | 1  | 93.97      | 95.22 | 93.77 | 91.41     | 93.82 | 94.56 | 94.88 |
|                | 5  | 94.34      | 94.91 | 93.72 | 93.48     | 95.06 | 95.37 | 95.46 |
|                | 10 | 94.41      | 94.58 | 93.82 | 93.85     | 95.27 | 95.50 | 95.56 |
|                | 15 | 94.44      | 94.30 | 93.82 | 94.23     | 95.51 | 95.66 | 95.53 |
|                | 20 | 94.56      | 94.22 | 93.82 | 94.47     | 95.50 | 95.58 | 95.47 |
|                | 25 | 94.60      | 94.17 | 93.82 | 94.57     | 95.53 | 95.62 | 95.23 |
|                | 30 | 94.60      | 94.12 | 93.82 | 94.68     | 95.57 | 95.57 | 95.18 |
| Hyperparameter |    | 1          | 2     | 3     | 0.01      | 0.05  | 0.1   | 0.2   |
|                |    | exponent   |       |       | gamma     |       |       |       |

ภาพที่ 9 การเปรียบเทียบการปรับแต่งค่าไฮเปอร์พารามิเตอร์ ด้วยอัลกอริทึม SMO

5.5 ผลการทดลองการจำแนกประเภทข้อมูลด้วยวิธี DT ด้วยอัลกอริทึม J48 ซึ่งเป็นอัลกอริทึมต้นไม้ตัดสินใจที่สามารถรองรับข้อมูลทั้งแบบไม่ต่อเนื่องและแบบต่อเนื่อง โดยเมื่อพิจารณาค่าความถูกต้อง จากการปรับค่าไฮเปอร์พารามิเตอร์ พบว่าค่าที่เหมาะสมซึ่งให้ค่าความถูกต้องสูงสุด ได้แก่ ค่า confidenceFactor เท่ากับ 0.25 และค่า minNumObj เท่ากับ 2 ผลลัพธ์ของการทดลองสามารถแสดงได้ดังภาพที่ 10

|                |      |       |       |                  |          |
|----------------|------|-------|-------|------------------|----------|
| minNumObj      | 1    | 94.05 | 94.02 | 93.73            | Accuracy |
|                | 2    | 94.17 | 94.03 | 93.80            |          |
|                | 3    | 94.12 | 93.95 | 93.56            |          |
|                | 4    | 94.01 | 93.96 | 93.83            |          |
|                | 5    | 94.02 | 93.86 | 93.66            |          |
| Hyperparameter | 0.25 | 0.50  | 0.75  | confidenceFactor |          |
|                |      |       |       |                  |          |

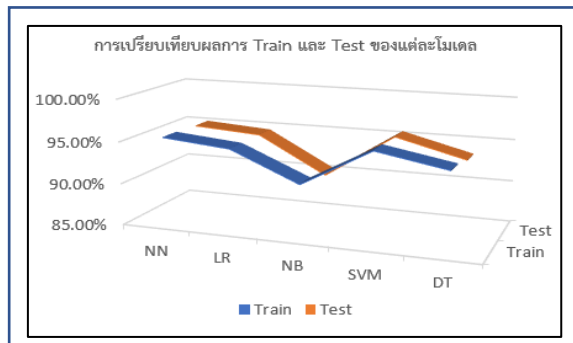
ภาพที่ 10 การเปรียบเทียบการปรับแต่งค่าไฮเปอร์พารามิเตอร์ ด้วยอัลกอริทึม J48

จากการประเมินประสิทธิภาพของแบบจำลองการจำแนกโดยใช้ค่าความถูกต้องเป็นตัวชี้วัด พบว่า ค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมซึ่งให้ค่าความถูกต้องสูงสุดของแต่ละแบบจำลองสามารถสรุปและแสดงได้ดังตารางที่ 6

ตารางที่ 6 ค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมมากที่สุดของแต่ละแบบจำลอง

| วิธีการเรียนรู้ของเครื่อง | ค่าไฮเปอร์พารามิเตอร์ที่เหมาะสม                             |
|---------------------------|-------------------------------------------------------------|
| NN                        | hidden layer= a, learning rate= 0.01                        |
| LR                        | maxIts= -1                                                  |
| NB                        | useKernelEstimator= FALSE, useSuperviseDiscretization= TRUE |
| SVM                       | Kernel = RBFkernel โดยที่ให้ค่า gamma = 0.1 และ C = 15.0    |
| DT                        | confidencefactor = 0.25, minNumObj = 2                      |

จากนั้น นำค่าไฮเปอร์พารามิเตอร์ของแต่ละอัลกอริทึมถูกนำมาใช้กับชุดข้อมูลสำหรับการฝึกสอนโมเดล และทดสอบโมเดล โดยสามารถแสดงผลได้ตามภาพที่ 11



ภาพที่ 11 การเปรียบเทียบค่าความถูกต้องของการฝึกสอน และการทดสอบโมเดล

จากผลการทดลองเพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องทั้ง 5 วิธี พบว่า แบบจำลองที่ใช้วิธี SVM ซึ่งดำเนินการด้วยอัลกอริทึม SMO มีประสิทธิภาพสูงสุดเมื่อพิจารณาจาก ตัวชี้วัดด้านความถูกต้อง โดยมีค่าความถูกต้อง (accuracy) เท่ากับ 95.55% ค่าความแม่นยำ (precision) เท่ากับ 95.60% ค่าความระลึก (recall) เท่ากับ 95.60% และค่าความถ่วงดุล (F1-score) เท่ากับ 95.50% โดยผลการเปรียบเทียบประสิทธิภาพของแบบจำลองทั้งหมดในภาพรวมแสดงไว้ดังตารางที่ 7

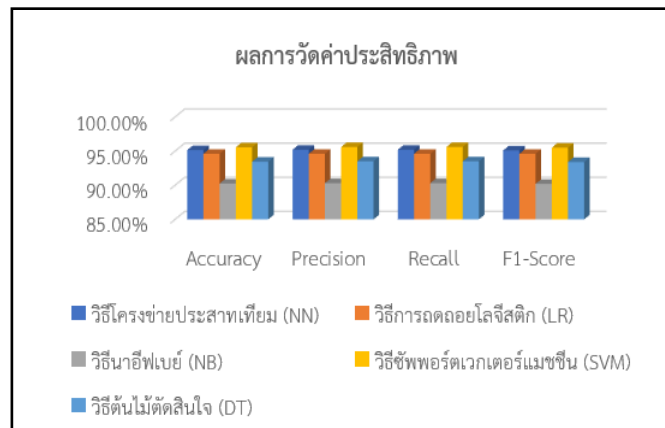
เมื่อพิจารณาผลการประเมินประสิทธิภาพเชิงลึกโดยใช้ confusion matrix ซึ่งแสดงรายละเอียดของค่าความแม่นยำ ความระลึก และค่าความถ่วงดุลของแต่ละแบบจำลองดังแสดงดังตารางที่ 8 พบว่า วิธี SVM ด้วยอัลกอริทึม SMO ให้ค่าตัวชี้วัดดังกล่าวสูงที่สุด แสดงให้เห็นว่าแบบจำลองมีความสามารถในการจำแนกยูอาร์แอลพิษซึ่งได้อย่างมีประสิทธิภาพ และมีเสถียรภาพสูงกว่าวิธีอื่น ดังนั้น ในภาพรวมของการจำลองการจำแนกยูอาร์แอลพิษซึ่งวิธี SVM จึงเป็นวิธีที่ให้ผลการจำแนกที่ดีที่สุด โดยแสดงกราฟการเปรียบเทียบผลลัพธ์ดังภาพที่ 12

ตารางที่ 7 ผลการวัดค่าประสิทธิภาพในการจำแนกของแบบจำลอง

| วิธีการเรียนรู้ของเครื่อง | ผลการวัดค่าประสิทธิภาพ |                |                |                |
|---------------------------|------------------------|----------------|----------------|----------------|
|                           | accuracy               | precision      | recall         | F1-score       |
| NN                        | 95.15 %                | 95.20 %        | 95.20 %        | 95.10 %        |
| LR                        | 94.60 %                | 94.60 %        | 94.60 %        | 94.60 %        |
| NB                        | 90.25 %                | 90.30 %        | 90.30 %        | 90.20 %        |
| <b>SVM</b>                | <b>95.55 %</b>         | <b>95.60 %</b> | <b>95.60 %</b> | <b>95.50 %</b> |
| DT                        | 93.45 %                | 93.50 %        | 93.50 %        | 93.40 %        |

ตารางที่ 8 การประเมินผลลัพธ์การทำนายโดยใช้ confusion matrix

| วิธีการเรียนรู้<br>ของเครื่อง | confusion matrix |            | ผลลัพธ์จากการทำนาย |          |
|-------------------------------|------------------|------------|--------------------|----------|
|                               |                  |            | legitimate         | phishing |
| NN                            | ผลลัพธ์จริง      | Legitimate | 941                | 59       |
|                               |                  | Phishing   | 38                 | 962      |
| LR                            | ผลลัพธ์จริง      | Legitimate | 947                | 53       |
|                               |                  | Phishing   | 55                 | 945      |
| NB                            | ผลลัพธ์จริง      | Legitimate | 913                | 87       |
|                               |                  | Phishing   | 108                | 892      |
| SVM                           | ผลลัพธ์จริง      | Legitimate | 956                | 44       |
|                               |                  | Phishing   | 45                 | 955      |
| DT                            | ผลลัพธ์จริง      | Legitimate | 938                | 62       |
|                               |                  | Phishing   | 69                 | 931      |

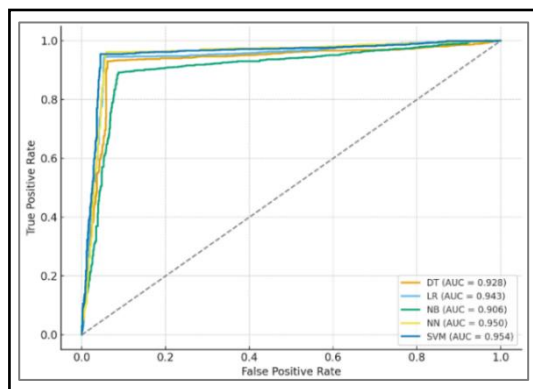


ภาพที่ 12 กราฟผลการวัดค่าประสิทธิภาพในการจำแนกของแบบจำลอง

5.6 ผลการวัดประสิทธิภาพด้วยเส้นโค้ง ROC (receiver operating characteristic curve) โดยเส้นโค้ง ROC เป็นกราฟที่ใช้ประเมินประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องสำหรับการจำแนกข้อมูลแบบสองคลาส โดยกราฟจะแสดงค่า อัตราการจำแนกผิดเป็นบวก (false positive rate: FPR) บนแกน X และค่า อัตราการจำแนกถูกเป็นบวก (true positive rate: TPR) บนแกน Y เพื่อใช้เปรียบเทียบความสามารถของแบบจำลองในการแยกแยะข้อมูลที่เป็นจริง และข้อมูลที่เป็นเท็จ กราฟที่มีเส้นโค้ง ROC ลู่เข้าใกล้มุมซ้ายบนของกราฟมากที่สุด สะท้อนให้เห็นว่าแบบจำลองนั้นมีประสิทธิภาพในการจำแนกข้อมูลสูงกว่าแบบจำลองอื่น (Sanjay, 2024)

โดยที่ค่า AUC (area under the curve) หมายถึงพื้นที่ใต้เส้นโค้ง ROC ซึ่งใช้เป็นตัวชี้วัดประสิทธิภาพโดยรวมของแบบจำลองการจำแนกประเภทในการแยกแยะระหว่างคลาสต่าง ๆ ผลการวิเคราะห์ในงานวิจัยนี้ แสดงให้เห็นว่าแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน (SVM) ด้วยอัลกอริทึม SMO มีเส้นโค้ง ROC ลู่เข้าใกล้มุมซ้ายบนมากที่สุด และมีค่า AUC สูงสุด เมื่อเปรียบเทียบกับแบบจำลองอื่น ได้แก่ โครงข่ายประสาทเทียม (NN) การถดถอยโลจิสติก (LR) นาอ็พเบย์ (NB) และต้นไม้ตัดสินใจ (DT) ซึ่งบ่งชี้ถึงประสิทธิภาพในการจำแนกยูอาร์แอลฟิชซิง

และยูอาร์แอลที่ถูกกฎหมายได้ดีที่สุด ขณะที่แบบจำลองนาอ์ฟเบส (NB) ให้ค่า AUC ต่ำที่สุด สะท้อนถึงประสิทธิภาพในการจำแนกข้อมูลต่ำกว่าแบบจำลองอื่น ผลการเปรียบเทียบดังกล่าว แสดงไว้ดังภาพที่ 13



ภาพที่ 13 กราฟเส้นโค้ง ROC เปรียบเทียบประสิทธิภาพการจำแนกของแต่ละโมเดล

## 6. สรุปผล และอภิปรายผล

งานวิจัยนี้สามารถพัฒนาแบบจำลองการเรียนรู้ของเครื่องสำหรับการวิเคราะห์ยูอาร์แอลฟิชซิงได้สำเร็จตามวัตถุประสงค์ที่กำหนดไว้ โดยใช้โปรแกรม weka ในการสร้าง และประเมินแบบจำลองการจำแนกประเภท ซึ่งแสดงให้เห็นว่าแบบจำลองที่พัฒนาขึ้นสามารถนำไปประยุกต์ใช้ในการตรวจจับเว็บไซต์ที่มีพฤติกรรมหลอกลวงได้จริง และมีศักยภาพในการพัฒนาต่อยอดเพื่อเพิ่มความแม่นยำ และลดความผิดพลาดในการตรวจจับในอนาคต

จากการประเมินประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องทั้ง 5 วิธี พบว่าแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนที่ใช้อัลกอริทึม SMO ให้ผลลัพธ์ที่ดีที่สุดในภาพรวม โดยมีค่าความถูกต้อง (accuracy) เท่ากับ 95.55% ค่าความแม่นยำ (precision) เท่ากับ 95.60% ค่าความระลึก (recall) เท่ากับ 95.60% และค่าความถ่วงดุล (F1-score) เท่ากับ 95.50% ซึ่งสะท้อนให้เห็นถึงความสามารถของแบบจำลองในการจำแนกยูอาร์แอลฟิชซิงได้อย่างมีประสิทธิภาพ และแม่นยำสูงกว่าวิธีอื่น โดยเฉพาะเมื่อเปรียบเทียบกับโครงข่ายประสาทเทียม (NN) ที่ใช้อัลกอริทึม MLP ซึ่งมีค่าความถูกต้องต่ำกว่าประมาณ 0.40% อย่างมีนัยสำคัญ โดยที่อัลกอริทึมที่มีประสิทธิภาพรองลงมาได้แก่ วิธีการถดถอยโลจิสติก มีค่าความถูกต้องเท่ากับ 94.60% วิธีต้นไม้ตัดสินใจมีค่าความถูกต้องเท่ากับ 93.45% และวิธีนาอ์ฟเบส มีค่าความถูกต้องต่ำที่สุด เท่ากับ 90.25%

เมื่อพิจารณาในเชิงการตีความผลลัพธ์ พบว่าวิธีซัพพอร์ตเวกเตอร์แมชชีนด้วยอัลกอริทึม SMO มีความเหมาะสมต่อปัญหาการจำแนกยูอาร์แอลฟิชซิง เนื่องจากสามารถสร้างเส้นแบ่งการจำแนกที่มีระยะขอบชัดเจน และรองรับข้อมูลที่มีมิติสูงได้อย่างมีประสิทธิภาพ โดยเฉพาะเมื่อใช้เคอร์เนลฟังก์ชันฐานรัศมีร่วมกับการปรับค่าไฮเปอร์พารามิเตอร์ด้วยวิธีการค้นหาแบบกริด ซึ่งช่วยให้แบบจำลองสามารถเรียนรู้รูปแบบความสัมพันธ์ที่ซับซ้อนของคุณลักษณะยูอาร์แอลได้ดีกว่าวิธีการอื่น

ผลการศึกษานี้มีความสอดคล้องกับงานวิจัยของ Rashid et al. (2020) และ Abutaha et al. (2021) ซึ่งรายงาน bahwaแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนให้ค่าความแม่นยำสูงในการตรวจจับเว็บไซต์ฟิชซิงเช่นเดียวกัน อย่างไรก็ตาม ความแตกต่างของงานวิจัยนี้อยู่ที่การประยุกต์ใช้

กระบวนการปรับค่าไฮเปอร์พารามิเตอร์อย่างเป็นระบบร่วมกับการใช้ชุดข้อมูลที่มีความสมดุล ซึ่งช่วยเพิ่มความเสถียรของผลลัพธ์ และลดความเอนเอียงของแบบจำลองในการจำแนกข้อมูล

แม้ว่างานวิจัยนี้จะแสดงให้เห็นถึงประสิทธิภาพของแบบจำลองที่พัฒนาขึ้น แต่ยังมีข้อจำกัดบางประการ ได้แก่ การใช้ข้อมูลยูอาร์แอลจากฐานข้อมูลสาธารณะ ซึ่งอาจยังไม่ครอบคลุมความหลากหลายของรูปแบบการโจมตีฟิชชิ่งในสภาพแวดล้อมจริง รวมถึงข้อจำกัดด้านเครื่องมือและฟังก์ชันของโปรแกรมที่ใช้ในการทดลอง ซึ่งอาจส่งผลต่อความสามารถในการวิเคราะห์เชิงลึกของแบบจำลอง

ดังนั้น งานวิจัยในอนาคตควรพิจารณาการขยายชุดข้อมูลให้มีความหลากหลายมากขึ้น รวมถึงการประยุกต์ใช้เทคนิคการคัดเลือกคุณลักษณะ (feature selection) และการพัฒนาโปรแกรมเพิ่มเติมให้มีความเฉพาะเจาะจงต่อบริบทการใช้งานจริง ซึ่งจะช่วยเพิ่มประสิทธิภาพความแม่นยำของแบบจำลอง และสนับสนุนการพัฒนาระบบตรวจจับยูอาร์แอลฟิชชิ่งแบบอัตโนมัติที่มีความน่าเชื่อถือมากยิ่งขึ้น

## 7. กิตติกรรมประกาศ

บทความวิจัยฉบับนี้ใช้เครื่องมือปัญญาประดิษฐ์เชิงกำเนิด (generative artificial intelligence) เพื่อช่วยปรับปรุงความชัดเจนของภาษาและสำนวนการเขียนเท่านั้น

## 8. เอกสารอ้างอิง

- นิสริวัฒน์ เจนศิริศักดิ์, ดารารัตน์ ทาสาจันทร์ และ ปวีณา วันชัย. (2568). การวิเคราะห์ความรู้สึกจากบทวิจารณ์ภาษาไทยของผู้บริโภคที่มีต่อสมาร์ตโฟน ในระดับมุมมอง. **วารสารวิทยาศาสตร์ลาดกระบัง** 34(1). 20-42.
- รัชนิวรรณ ไพศาลรเกียรติ. (2564). การเปรียบเทียบตัวแบบการถดถอยลอจิสติกและเทคนิคเหมืองข้อมูลสำหรับพยากรณ์การเป็นโรคเบาหวาน. วิทยานิพนธ์ปริญญาโทมหาบัณฑิต ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์. มหาวิทยาลัยนเรศวร, ประเทศไทย.
- Abutaha, M., Ababneh, M., Mahmoud, K., & Baddar, S. A.-H. (2021). **URL Phishing Detection using Machine Learning Techniques based on URLs Lexical Analysis**. 2021 12th International Conference on Information and Communication Systems (ICICS), 147–152.  
<https://doi.org/10.1109/ICICS52457.2021.9464539>.
- Ademola P.A. and Boniface K. (2021). **Phishing Attack in Communication Networks is exposed using a Multi-Stage Machine Learning Approach**. ECTI TRANSACTIONS ON COMPUTER AND INFORMATION TECHNOLOGY.  
<https://doi.org/10.37936/ecti-cit.2021153.240565>.
- Ahammad, S. H., Kale, S. D., Upadhye, G. D., Pande, S. D., Babu, E. V., Dhumane, A. V., & Bahadur, Mr. D. K. J. (2022). **Phishing URL detection using machine**

- learning methods.** *Advances in Engineering Software*, 173, 103288.  
<https://doi.org/10.1016/j.advengsoft.2022.103288>.
- Aljofey, A., Jiang, Q., Qu, Q., Huang, M., & Niyigena, J.-P. (2020). **An Effective Phishing Detection Model Based on Character Level Convolutional Neural Network from URL.** *Electronics*, 9(9), 1514. <https://doi.org/10.3390/electronics9091514>.
- Anggreani, D., Hamdani, & Nurmisba. (2024). **Grid Search Hyperparameter Analysis in Optimizing The Decision Tree Method for Diabetes Prediction.** *Indonesian Journal of Data and Science*, 5(3), 190–197.  
<https://doi.org/10.56705/ijodas.v5i3.190>.
- Ashok, A., Rathis, D., Raghavendra, R., & Umadevi, V. (2024). **A Comparative Analysis of Traditional Machine Learning, Deep Learning and Boosting Algorithms on Phishing URL Detection.** 2024 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI), 1–6.  
<https://doi.org/10.1109/CVMI61877.2024.10782525>.
- Chaiyaphop Jamjumrat. (2022), **รู้จักกับ Decision Tree มันคือต้นไม้อะไร ทำงานอย่างไร?** สืบค้นจาก <https://www.borntodev.com/2022/09/15/รู้จักกับ-decision-tree>.
- Elgeldawi, E., Sayed, A., Galal, A.R. and Zaki, A.M. (2021) **Hyperparameter Tuning for Machine Learning Algorithms Used for Arabic Sentiment Analysis.** *Informatics*, 8, Article 79. <https://doi.org/10.3390/informatics8040079>.
- Ghalechyan, H., Israyelyan, E., Arakelyan, A., Hovhannisyan, G., & Davtyan, A. (2024). **Phishing URL detection with neural networks: An empirical study.** *Scientific Reports*, 14(1), 25134. <https://doi.org/10.1038/s41598-024-74725-6>.
- Hannousse, Abdelhakim, Yahiouche, Salima. (2021). **Web page phishing detection.** Retrieved from <https://www.kaggle.com/datasets/shashwatwork/web-page-phishing-detection-dataset/data>.
- Haq, Q. E. U., Faheem, M. H., & Ahmad, I. (2024). **Detecting Phishing URLs Based on a Deep Learning Approach to Prevent Cyber-Attacks.** *Applied Sciences*, 14(22), 10086. <https://doi.org/10.3390/app142210086>.
- Justus, A. I., Durodola, O., Alade, O., J Awotunde, O., T Olanrewaju, A., Falana, O., Ogungbire, A., Osinuga, A., Ogunbiyi, D., Ifeanyi, A., E Odezuligbo, I., & E Edu, O. (2024). **Hyperparameter Tuning in Machine Learning: A Comprehensive Review.** *Journal of Engineering Research and Reports*, 26(6), 388–395.  
<https://doi.org/10.9734/jerr/2024/v26i61188>.
- Pasith Thanapatpisarn. (2022), **Decision Tree โมเดลต้นไม้ กับความเรียงมากในการตัดสินใจ!!! Part1**, สืบค้นจาก <https://datascihaeng.medium.com/decision-tree-part01-47ef24539fba>.

- Rashid, J., Mahmood, T., Nisar, M. W., & Nazir, T. (2020). **Phishing Detection Using Machine Learning Technique**. 2020 First International Conference of Smart Systems and Emerging Technologies (SMARTTECH), 43–46.  
<https://doi.org/10.1109/SMART-TECH49988.2020.00026>.
- Safi, A., & Singh, S. (2023). **A systematic literature review on phishing website detection techniques**. Journal of King Saud University - Computer and Information Sciences, 35(2), 590–611.  
<https://doi.org/10.1016/j.jksuci.2023.01.004>.
- Salloum, S., Gaber, T., Vadera, S., & Shaalan, K. (2021). **Phishing Email Detection Using Natural Language Processing Techniques: A Literature Survey**. Procedia Computer Science, 189, 19–28.  
<https://doi.org/10.1016/j.procs.2021.05.077>.
- Sanjay Dutta. (2024). **Understanding the ROC Curve: When and How to Use It in Binary Classification**. Retrieved from [https://medium.com/@sanjay\\_dutta/understanding-the-roc-curve-when-and-how-to-use-it-in-binary-classification-724b97f641f4](https://medium.com/@sanjay_dutta/understanding-the-roc-curve-when-and-how-to-use-it-in-binary-classification-724b97f641f4).
- Stephen Oladele. (2024). **Top 12 Dimensionality Reduction Techniques for Machine Learning**. Retrieve from <https://encord.com/blog/dimentionality-reduction-techniques-machine-learning>.