การวิเคราะห์แบบจำลองข้อมูลเพื่อจำแนกโดเมน DGA และ Legitimate จากผู้ประสงค์ร้ายในการเข้าใช้อินเทอร์เน็ต

Analysis of Data Model to Distinguish DGA and Legitimate Domains from Malicious Person Accessing the Internet

สุกฤษฎิ์ สรหงษ์*
Sugrid Sorrahong
อรรถพล ป้อมสถิตย์**
Auttapon Pomsathit
บุญเรือง เกิดอรุณเดช***
Boonruang Kerdaroondej

Received : February 17, 2022
Revised : April 7, 2022
Accepted : July 25, 2022

บทคัดย่อ

งานวิจัยฉบับนี้นำเสนอแนวทางในการนำวิธีการเรียนรู้ของเครื่องจักรแบบมีผู้สอน (Supervised Machine Learning) มาใช้ในการวิเคราะห์ และจำแนกประเภท (Classification) ระหว่างชื่อโดเมนที่ถูกต้อง (Legitimate Domain Name) และชื่อโดเมนที่ถูกสร้างขึ้นจากอัลกอริทึมสำหรับสร้างโดเมน (Domain Generation Algorithms: DGA) โดยนำรูปแบบในการตัดสินของมนุษย์มากำหนดเป็นแอตทริบิว (Attribute) ที่ใช้ในการวิเคราะห์ เช่น ความยาวของชื่อโดเมน จำนวนตัวอักษรและตัวเลขที่ถูกนำมาประกอบในชื่อโดเมน (Domain Name) จำนวนคำที่มีความหมายที่อยู่ในชื่อโดเมน และจำนวนคำที่สามารถอ่านออกเสียงได้ที่อยู่ใน

*นักศึกษาหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาการจัดการความมั่นคงปลอดภัยไซเบอร์ วิทยาลัยนวัตกรรมดิจิทัล เทคโนโลยี มหาวิทยาลัยรังสิต
Master of Science Program in Cybersecurity Management College of Digital Innovation Technology, Rangsit University
**อาจารย์หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาการจัดการความมั่นคงปลอดภัยไซเบอร์ วิทยาลัยนวัตกรรมดิจิทัล เทคโนโลยี มหาวิทยาลัยรังสิต
Lecturer in Master of Science Program in Cybersecurity Management College of Digital Innovation Technology, Rangsit University
***ผู้เชี่ยวชาญด้านยุทธศาสตร์การพัฒนาพฤติกรรม และระบบสารสนเทศ, กรมสนับสนุนบริการสุขภาพ, กระทรวงสาธารณสุข
Consultant and Expert in Strategy, Behavior Modification, Information Technology, Department of Service Support, Ministry of Public Health

ชื่อโดเมน เป็นต้น โดยนำตัวอย่างชื่อโดเมนที่ถูกต้อง จำนวน 50,000 รายการ และชื่อโดเมนที่ถูกสร้างขึ้นจากอัลกอริทึม สำหรับสร้างโดเมน จำนวน 50,000 รายการ รวมทั้งสิ้น 100,000 รายการ เป็นข้อมูลตั้งต้นเพื่อป้อนให้กับแบบจำลองต้นไม้การตัดสินใจแบบจำแนกประเภทและแบบถดถอย (Classification and Regression Tree: CART) แบ่งเป็นชุดข้อมูลสำหรับเรียนรู้ (Training Data) จำนวน 70% และชุดข้อมูลสำหรับทดสอบ (Testing Data) จำนวน 30% โดยใช้ภาษาไพทอน (Python) และชุดคำสั่งที่สำคัญ (Library) เป็นเครื่องมือในการวิเคราะห์ เมื่อวัดประสิทธิภาพของแบบจำลองด้วยชุดข้อมูลสำหรับทดสอบ หลังจากปรับปรุงประสิทธิภาพของแบบจำลองด้วยวิธีการตัดกิ่งในขณะเรียนรู้ (Pre Pruning) โดยใช้เมทริกซ์ความสับสน (Confusion Matrix) พบว่าแบบจำลองสามารถจำแนกประเภทระหว่างระหว่างชื่อโดเมนที่ถูกต้อง และ ชื่อโดเมนที่ถูกสร้างขึ้นจากอัลกอริทึมสำหรับสร้างโดเมนได้อย่างมีประสิทธิภาพ โดยให้ค่าความถูกต้อง (Accuracy) 97.25% ความแม่นยำ (Precision) 96.25% ค่าการระลึกได้ (Recall) 97.25% และค่าเฉลี่ย Harmonic Mean (F1 Score) 96.75%

คำสำคัญ : การเรียนรู้ของเครื่องจักรแบบมีผู้สอน / ชื่อโดเมน / อัลกอริทึมสำหรับสร้างโดเมน / ต้นไม้การตัดสินใจแบบจำแนกประเภทและแบบถดถอย

ABSTRACT

This research proposed an approach to applying Supervised Machine Learning methods to the analysis and classification between legitimate domain names and domain names generated by Domain Generation Algorithms (DGA). By human decision-making methods to define attributes in domain names analysis. For instance, the length of a domain name, the number of letters and numbers that are components of a domain name, the number of meaningful words in a domain name, and the number of pronounceable words in a domain name. The population of huge domain names consists of 50,000 legitimate domain names and 50,000 DGA. Then are split into 70% training datasets and 30% testing datasets before being fed into the classification and regression tree model (CART) using Python and Libraries. After improving the efficiency of the model with the pre-pruning method and evaluating the performance of the model with the confusion matrix, the decision tree model classifies between legitimate domain names and DGA more efficiently, which provides accuracy is 97.25%, precision is 96.25%, recall is 97.25%, and F1 score is 96.75%.

Keywords : Supervised Machine Learning / Domain Name / Domain Generation Algorithms (DGA) / Classification and Regression Tree (CART)

บทนำ

ปัจจุบัน ระบบเทคโนโลยีสารสนเทศเป็นระบบที่มีความสำคัญต่อการพัฒนาของประเทศในทุกๆ ด้าน ไม่ว่าจะเป็นด้านความมั่นคง เศรษฐกิจ และสังคม มีความสำคัญต่อการขับเคลื่อนขององค์กรในหลายมิติให้สำเร็จลุล่วงได้อย่างรวดเร็วและมีประสิทธิภาพ และในขณะเดียวกันมีผู้ไม่ประสงค์ดีได้นำระบบเทคโนโลยีสารสนเทศมาใช้เป็นเครื่องมือสำหรับโจมตีระบบอื่นๆ ส่งผลให้ไม่สามารถทำงานได้อย่างถูกต้อง (Integrity) สูญเสียข้อมูลที่เป็นความลับ (Confidentiality) หรือแม้กระทั่งถูกระงับการให้บริการ (Availability) เหล่านี้ล้วนเป็นภัยคุกคามที่ส่งผลกระทบโดยตรงต่อระบบเทคโนโลยีสารสนเทศขององค์กรทั้งสิ้น ขณะนี้มีภัยคุกคามในรูปแบบต่างๆ เกิดขึ้นเป็นจำนวนมาก และมีแนวโน้มที่จะทวีความรุนแรงขึ้นอย่างต่อเนื่อง จากข้อมูลการจัดอันดับภัยคุกคามทางไซเบอร์ที่แสดงในเอกสาร ENISA Threat Landscape 15 Top Threats in 2020 (The European Union Agency for Cybersecurity, 2020) เผยแพร่โดยหน่วยงาน The European Union Agency for Cybersecurity (ENISA) ซึ่งมีหน้าที่รับผิดชอบเกี่ยวกับงานด้านความมั่นคงปลอดภัยไซเบอร์ในสหภาพยุโรป พบว่าภัยคุกคามทางไซเบอร์ 3 อันดับแรก คือ 1. มัลแวร์ (Malware) 2. การโจมตีเว็บไซต์ (Web-based attacks) และ 3. ฟิชชิ่ง (Phishing) โดยเฉพาะอย่างยิ่งมัลแวร์ ซึ่งถูกจัดอันดับให้เป็นภัยคุกคามทางไซเบอร์อันดับที่ 1 นั้น มักถูกพบว่ามีการนำอัลกอริทึมสำหรับสร้างโดเมน (Domain Generation Algorithms: DGA) มาใช้เป็นเครื่องมือในการโจมตีระบบเทคโนโลยีสารสนเทศต่างๆ ร่วมกับมัลแวร์ โดยมีวัตถุประสงค์เพื่อใช้อัลกอริทึมสำหรับสร้างโดเมนในการปิดบังช่องทางที่แท้จริงที่ผู้ไม่ประสงค์ดีใช้ในการรับและส่งข้อมูล รวมถึงใช้ในการโจมตี ด้วยเหตุนี้ อัลกอริทึมสำหรับสร้างโดเมน จึงเปรียบเสมือนภัยคุกคามทางไซเบอร์รูปแบบหนึ่งที่องค์กรควรให้ความสำคัญในการตรวจสอบ และเตรียมการรับมือ เพื่อลดความเสี่ยงต่างๆ ที่อาจเกิดขึ้นกับข้อมูล และระบบเทคโนโลยีสารสนเทศที่มีความสำคัญขององค์กร

อัลกอริทึมสำหรับสร้างโดเมน เป็นขั้นตอนหรือวิธีการหนึ่งที่ผู้ไม่ประสงค์ดีมักนำมาใช้งานร่วมกับมัลแวร์เพื่อซ่อนพรางชื่อโดเมน (Domain Name) ที่แท้จริง ที่ผู้ไม่ประสงค์ดีได้จดทะเบียนไว้สำหรับใช้เป็นช่องทางในการติดต่อย้อนกลับ (Call back) ไปยังเครื่องเซิร์ฟเวอร์ที่ใช้สำหรับควบคุมและสั่งการ (Command and Control: C2) (Chowdhury, 2019) โดยการสร้างชื่อโดเมนออกมาอย่างต่อเนื่องเป็นจำนวนมาก ด้วยวิธีการสุ่มอย่างมีรูปแบบ ซึ่งล้วนเป็นชื่อโดเมนที่ไม่สามารถจดจำ หรือทำความเข้าใจได้โดยง่าย เช่น nlik88f8hokpv81g.com และ 1tobh33u4gyfuu6o7mz12j7v08.com เป็นต้น หลังจากที่ระบบหรือเครื่องของเป้าหมายติดเชื้อมัลแวร์ จะส่งข้อมูลเพื่อใช้สำหรับการติดต่อย้อนกลับระหว่างระบบหรือเครื่องเป้าหมายที่ติดเชื้อ กับเครื่องเซิร์ฟเวอร์ที่ใช้สำหรับควบคุมและสั่งการออกไปตามรายชื่อโดเมนที่ถูกสร้างขึ้นทั้งหมด ซึ่งผู้ไม่ประสงค์ดีได้นำรายชื่อโดเมนที่ถูกสร้างขึ้นด้วยวิธีการดังกล่าวบางส่วนไปจดทะเบียนเพื่อรับหมายเลขที่อยู่บนระบบเครือข่าย (IP Address) รอไว้ก่อนหน้า โดยนำไปใช้งานร่วมกับเครื่องเซิร์ฟเวอร์ที่ใช้สำหรับควบคุมและสั่งการ เพื่อรอการติดต่อย้อนกลับจากระบบหรือเครื่องเป้าหมายที่ติดเชื้อ หากมีรายชื่อโดเมน หรือหมายเลขที่อยู่บนระบบเครือข่ายรายการใดถูกตรวจพบ และถูกปิดกั้นการเชื่อมต่อ (Blocked) ผู้ไม่ประสงค์ดีสามารถนำรายชื่อโดเมนอื่นที่ได้จดทะเบียนไว้มาใช้ในการโจมตีต่อเนื่องได้ทันที จะเห็นได้ว่ากระบวนการดังกล่าวสร้างความ

ยากลำบากในการตรวจสอบเพื่อหารายชื่อโดเมนที่ใช้ในการโจมตีจริงเพียงไม่กี่รายการ จากรายชื่อโดเมนที่ถูกสร้างขึ้นมาเป็นจำนวนมากนี้ ส่งผลให้ผู้ดูแลระบบไม่สามารถป้องกัน หรือยับยั้งการเชื่อมต่อที่เป็นภัยคุกคามนี้ได้อย่างทันท่วงที

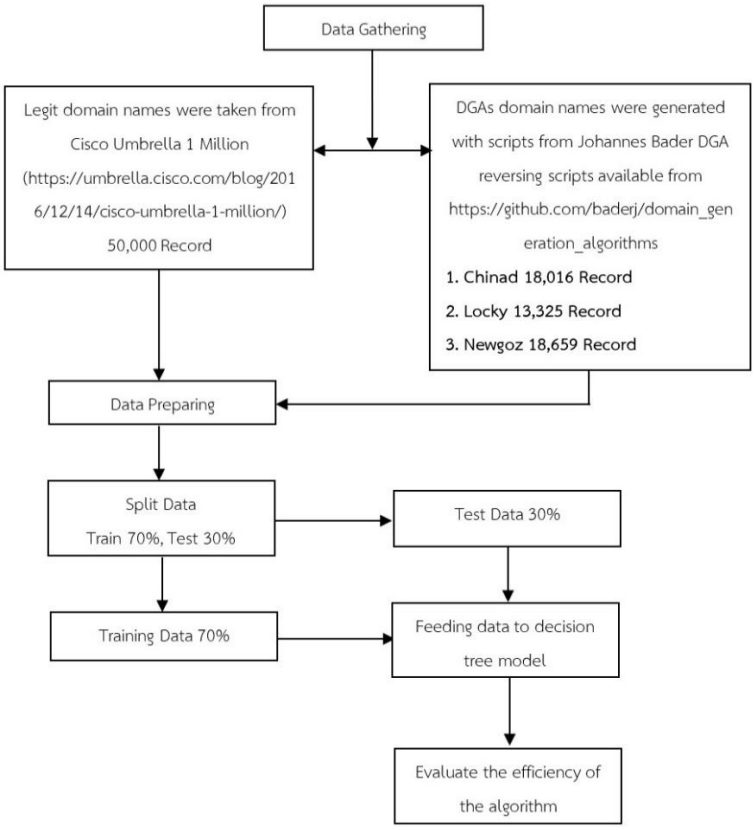
งานวิจัยฉบับนี้นำเสนอเกี่ยวกับแนวทางในการนำวิธีการเรียนรู้ของเครื่องจักรแบบมีผู้สอน (Supervised Machine Learning) (G.P., R., S., & Gladston, 2020) มาประยุกต์ใช้ในการวิเคราะห์ และจัดหมวดหมู่ (Classification) ชื่อโดเมน โดยนำรูปแบบในการตัดสินใจแบบมนุษย์มาป้อนให้กับเครื่องจักรได้นำไปเรียนรู้เพื่อหาความสัมพันธ์ จดจำรูปแบบ และลักษณะของชื่อโดเมนในแต่ละประเภท เช่น ความยาวของชื่อโดเมน จำนวนตัวอักษรและตัวเลขที่ถูกนำมาประกอบในชื่อโดเมน ความหมายของคำในชื่อโดเมน และความสามารถในการอ่านออกเสียงของคำในชื่อโดเมน โดยการนำตัวอย่างชื่อโดเมนที่อยู่ในรูปแบบต่าง ๆ ซึ่งประกอบด้วยชื่อโดเมนที่ถูกต้อง (Legitimate Domain Name) และชื่อโดเมนที่ถูกสร้างขึ้นมาจากอัลกอริทึมสำหรับสร้างโดเมนเป็นข้อมูลในการวิเคราะห์ โดยใช้เทคนิคต้นไม้ตัดสินใจ (Decision Tree) ในการประมวลผลเพื่อแสดงให้เห็นถึงลักษณะหรือรูปแบบที่ใช้ในการจัดหมวดหมู่หรือแยกประเภท รวมไปถึงการทำนายรูปแบบ (Prediction) ระหว่างชื่อโดเมนที่ถูกต้อง และชื่อโดเมนที่ถูกสร้างขึ้นจากอัลกอริทึมสำหรับสร้างโดเมน

วิธีดำเนินการวิจัย

การดำเนินการวิจัยประกอบด้วยขั้นตอนที่สำคัญ จำนวนทั้งสิ้น 5 ขั้นตอน ประกอบด้วย

1. การเก็บรวบรวมข้อมูล
2. การจัดเตรียมข้อมูลให้อยู่ในรูปแบบที่ถูกต้อง
3. การแบ่งข้อมูลเพื่อใช้ในการวิเคราะห์
4. การสร้างแบบจำลองเพื่อจำแนกประเภทโดยใช้กระบวนการต้นไม้ตัดสินใจแบบจำแนกประเภทและแบบถดถอย (Classification and Regression Tree: CART) เพื่อใช้ในการวิเคราะห์ข้อมูล
5. การประเมินประสิทธิภาพของแบบจำลอง

จากขั้นตอนทั้งหมดสามารถนำมาสรุปเป็นแผนภาพแสดงขั้นตอนในการดำเนินการวิจัยได้ดังนี้



ภาพที่ 1 แสดงผังขั้นตอนการดำเนินการวิจัย

1. การเก็บรวบรวมข้อมูล

ข้อมูลที่ใช้สำหรับการดำเนินการวิจัยในครั้งนี้ประกอบด้วย รายชื่อโดเมนที่ถูกต้อง (Legitimate Domain Name) จากโครงการ Cisco Umbrella 1 Million (Hubbard, 2016) มาใช้ในการวิจัย โดยข้อมูลดังกล่าวประกอบด้วยรายชื่อโดเมน (Domain Name) ที่ได้รับความนิยมสามารถพบเห็นได้ทั่วไปในขณะที่ท่องเว็บไซต์ จำนวน 1,000,000 รายชื่อ ซึ่งผู้ดำเนินการวิจัยได้ทำการปรับลดรายชื่อโดเมนสำหรับนำมาใช้ในการวิจัยลงเหลือเพียง 50,000 รายชื่อ และชื่อโดเมนที่ถูกสร้างขึ้นมาจากอัลกอริทึมสำหรับสร้างโดเมน (Domain Generation Algorithms: DGA) จากฐานข้อมูล GitHub โดย Johannes Bader ในหัวข้อ “domain_generation_algorithms” (Bader, 2015) ซึ่งมีเนื้อหาเกี่ยวกับผลลัพธ์ที่ได้จากการทำกระบวนการย้อนกลับ กับรายชื่อโดเมนที่ถูกสร้างขึ้นมาจากอัลกอริทึมสำหรับสร้างโดเมน จำนวน 47 รูปแบบ โดยผู้ดำเนินการวิจัยได้เลือกข้อมูลชื่อโดเมนที่ถูกสร้างขึ้นมาจากอัลกอริทึมสำหรับสร้างโดเมนที่มักถูกพบว่ามีการนำมาใช้งานร่วมกันอยู่บ่อยครั้ง เพื่อนำมาใช้ในการวิจัย จำนวน 3 รูปแบบ ได้แก่ Chinad จำนวน 18,016 รายการ Locky จำนวน 13,325 รายการ และ Newgoz จำนวน 18,659 รายการ รวมทั้งสิ้น 100,000 รายการ

2. การจัดเตรียมข้อมูลให้อยู่ในรูปแบบที่ถูกต้อง (กุญแจสำคัญ)

การที่เครื่องคอมพิวเตอร์จะสามารถนำข้อมูลไปประมวลผลได้ โดยนำรูปแบบในการตัดสินใจของมนุษย์มาป้อนให้กับเครื่องคอมพิวเตอร์ได้เรียนรู้ และนำไปใช้ในการวิเคราะห์นั้น ข้อมูลจะต้องถูกแปลงให้อยู่ในรูปแบบที่เครื่องคอมพิวเตอร์สามารถเข้าใจได้เสียก่อน ซึ่งขั้นตอนนี้ถือเป็นกุญแจที่สำคัญที่ใช้ในการกำหนดคุณลักษณะของแอตทริบิว (Attribute) เพื่อใช้เป็นข้อมูลสำหรับป้อนให้กับแบบจำลองในการจำแนกประเภทของอัลกอริทึมสำหรับสร้างโดเมน ประกอบด้วย 2.1 การกำหนดชื่อแอตทริบิวและค่าที่ใช้ในการประมวลผล และ 2.2 การทำความสะอาดข้อมูล (Data Cleaning)

2.1 การกำหนดชื่อแอตทริบิวและค่าที่ใช้ในการประมวลผล โดยการนำชุดคำสั่งที่สำคัญ (Library) ในภาษาไพธอน (Python) เช่น 1) Tldextract 2) Nltk_word 3) Pronouncing และ 4) Wordsegment มาใช้การประมวลผลเพื่อทำการกำหนดค่าให้กับแอตทริบิวต่างๆ จำนวนทั้งสิ้น 12 ตัว ได้แก่ subdomain, maindomain, length, word_length, num_length, pronouncing, pronounce_elements, can_pronounce, in_dict, word_elements, elements และ elements_indict โดยขอยกตัวอย่างชื่อโดเมน “www.flightradar24.com” เพื่อใช้สำหรับประกอบการทำความเข้าใจการใช้งานชุดคำสั่ง เพื่อกำหนดค่าให้กับแอตทริบิวที่สำคัญ ดังนี้

1) Tldextract ทำหน้าที่วิเคราะห์ชื่อโดเมน เพื่อทำการแยกโดเมนในระดับสูงสุด (Top Level Domain: TLD) ออกจากโดเมนย่อย (Subdomain) และส่วนต่อท้าย (Suffix) จากตัวอย่างข้างต้นเมื่อผ่านการประมวลผลโดยการใช้ Tldextract Library (Kurkowski, 2020) จะได้ผลเป็น Subdomain = ‘www’, Domain = ‘flightradar24’, Suffix = ‘com’ ดังนั้นแอตทริบิว subdomain จะมีค่า = ‘www’ และแอตทริบิว maindomain จะมีค่า = ‘flightradar 24’ ส่วนแอตทริบิว length หาค่าได้จากการนับจำนวนตัวอักษรและตัวเลขของแอตทริบิว maindomain รวมกัน = 13 ตัว แอตทริบิว word_length หาค่าได้จากการนับค่าจำนวนตัวอักษรของแอตทริบิว maindomain มีค่า = 11 ตัว และแอตทริบิว num_length หาค่าได้จากการนับค่าจำนวนตัวเลขของแอตทริบิว maindomain มีค่า = 2 ตัว

2) Pronouncing ทำหน้าที่วิเคราะห์ชื่อโดเมน ในที่นี้คือแอตทริบิว maindomain เพื่อตรวจสอบความสามารถในการอ่านออกเสียงของคำที่ถูกนำมาตั้งเป็นชื่อโดเมนว่าเป็นคำที่สามารถอ่านออกเสียงได้หรือไม่ จากตัวอย่างข้างต้นเมื่อผ่านการประมวลผลโดยการใช้ Pronouncing Library (Parrish, 2015) จะได้ผลเป็น [['F L AY1 T'], ['R EY1 D AA2 R'], []] ดังนั้นแอตทริบิว pronouncing จะมีค่า = [['F L AY1 T'], ['R EY1 D AA2 R'], []] จากนั้นแอตทริบิว pronounce_elements จะเก็บค่าที่ได้จากการนับจำนวนสมาชิกทั้งหมดที่ pronouncing เก็บไว้ = 3 ตัว และแอตทริบิว can_pronounce จะเก็บเฉพาะค่าของจำนวนสมาชิกที่สามารถอ่านออกเสียงจริงๆ เท่านั้น ซึ่งมีจำนวน = 2 ตัว

3) Nltk_word ทำหน้าที่วิเคราะห์ชื่อโดเมน เพื่อทำการตรวจสอบชื่อโดเมนหลัก (maindomain) ที่ยังไม่ผ่านการสกัดคำไปตรวจสอบกับฐานข้อมูลพจนานุกรม (ภาษาอังกฤษ) ที่บรรจุอยู่ใน

nltk.corpus โดยการใช้ nltk_word Library (Bird, Klein, & Loper, 2009, pp.261-289) ในการตรวจสอบจากตัวอย่างข้างต้นจะเห็นได้ว่า flightradar24 เป็นคำที่ไม่มีอยู่ในพจนานุกรม ดังนั้น in_dict จะมีค่า = 0

4) Wordsegment ทำหน้าที่สกัดคำออกมาจากชื่อโดเมนหลัก จากตัวอย่างข้างต้นจะเห็นได้ว่า flightradar24 หลังจากผ่านการประมวลผลโดยการใช้ wordsegment Library (Jenks, 2018) จะได้ผลเป็น ['flight', 'radar', '24'] ดังนั้นตัวแปร word_element จะมีค่า = ['flight', 'radar', '24'] โดยแอตทริบิว elements จะเก็บค่าที่ได้จากการนับจำนวนสมาชิกทั้งหมดที่ word_element ซึ่งมีค่า = 3 และแอตทริบิว elements_indict จะเก็บค่าจำนวนคำที่สามารถสกัดออกมาจากชื่อโดเมนหลัก และเป็นคำที่มีอยู่ในพจนานุกรม ซึ่งมีค่า = 2

เมื่อดำเนินการกำหนดชื่อแอตทริบิวและค่าที่ใช้ในการประมวลผลเสร็จสิ้นแล้วจะได้ไฟล์นามสกุล ".CSV" ดังแสดงตัวอย่างในภาพที่ 2

class_type	domain	subclass_type	subdomain	maindomain	length	word_length	num_length	pronouncing	pronounce_elements	can_pronounce	in_dict	word_elements	elements	elements_indict
legit	google.com	legit	www	google	6	6	0	[['G UW1 G AH0 L']]	1	1	0	['google']	1	0
legit	www.google.com	legit	www	google	6	6	0	[['G UW1 G AH0 L']]	1	1	0	['google']	1	0
legit	microsoft.com	legit	www	microsoft	9	9	0	[['M AY1 K R OW2 S AO1 F T']]	1	1	0	['microsoft']	1	0
dga	3jep1yaj1qxty3zizw13h3lmm.org	newgoz	www	3jep1yaj1qxty3zizw13h3lmm	25	18	7	[[], ['M']]	2	1	0	['3jep1yaj1qxty3zizw13h3l', 'mm']	2	0
dga	bz2fi81j9u5g97ibjymsezz1b.com	newgoz	www	bz2fi81j9u5g97ibjymsezz1b	25	17	8	[[], ['B IY1']]	2	1	0	['bz2fi81j9u5g97ibjymsezz1', 'b']	2	1
dga	1f9fxz41eo9bn3h5q4fp14vqxus.net	newgoz	www	1f9fxz41eo9bn3h5q4fp14vqxus	27	17	10	[[], ['EH1 K S'], ['AH1 S', 'Y UW2	3	2	0	['1f9fxz41eo9bn3h5q4fp14vq', 'x', 'us']	3	2
dga	5h9vh5oqp2699oi2.ru	chinad	www	5h9vh5oqp2699oi2	16	8	8	[[]]	1	0	0	['5h9vh5oqp2699oi2']	1	0
dga	you0krwoykp75dmg.biz	chinad	www	you0krwoykp75dmg	16	13	3	[['Y UW1'], []]	2	1	0	['you', '0krwoykp75dmg']	2	1
dga	7panmowwofbyytfe.net	chinad	www	7panmowwofbyytfe	16	15	1	[[]]	1	0	0	['7panmowwofbyytfe']	1	0
dga	bosevbyjjvvlno.xyz	locky	www	bosevbyjjvvlno	14	14	0	[[]]	1	0	0	['bosevbyjjvvlno']	1	0
dga	xmlmwax.click	locky	www	xmlmwax	7	7	0	[['EH2 K S EH2 M EH1 L'], []]	2	1	0	['xml', 'mwax']	2	0
dga	akuuwmmu.ru	locky	www	akuuwmmu	8	8	0	[[], [], []]	3	0	0	['aku', 'uw', 'mmu']	3	1

ภาพที่ 2 แสดงตัวอย่างข้อมูลหลังจากผ่านการประมวล เพื่อระบุค่าให้กับแอตทริบิว (Attribute) เสร็จสิ้นแล้ว

2.2 การทำความสะอาดข้อมูล (Data Cleaning) โดยการนำข้อมูลที่ผ่านกระบวนการกำหนดชื่อแอตทริบิวและค่าที่ใช้ในการประมวลผลเสร็จสิ้นแล้วมาตรวจสอบเพื่อกำจัดค่ามลทินหรือค่าที่ไม่พึงประสงค์ออกจากชุดข้อมูล ซึ่งอาจส่งผลให้เกิดข้อผิดพลาดในระหว่างการประมวลผล หรืออาจทำให้ได้ผลลัพธ์ที่ไม่ถูกต้อง และมีความคลาดเคลื่อน จากการประมวลผลข้อมูลดังกล่าวพบว่าข้อมูลที่ไม่สามารถนำมาประมวลผลเพื่อระบุค่าได้ จำนวน 13 รายการ ผู้ดำเนินการวิจัยจึงได้ทำความสะอาดข้อมูลด้วยวิธีการลบ (Delete) หรือนำกลุ่มข้อมูลที่มีปัญหา จำนวน 13 รายการ ออกจากชุดข้อมูลที่จะใช้ในการวิเคราะห์คิดเป็นอัตราส่วนโดยประมาณร้อยละ 0.013 ซึ่งถือเป็นกลุ่มข้อมูลที่มีขนาดเล็กมาก ไม่ส่งผลกระทบใดๆ ต่อกลุ่มข้อมูลส่วนใหญ่ซึ่งยังคงมีค่าใกล้เคียง 100,000 รายการอยู่ โดยคงเหลือข้อมูลที่จะนำไปใช้ในกระบวนการต่อไปทั้งสิ้น จำนวน 99,987 รายการ

3. การแบ่งข้อมูลเพื่อใช้ในการวิเคราะห์

การแบ่งข้อมูลเพื่อใช้สำหรับการเรียนรู้และการทดสอบ คือ การนำข้อมูลที่ได้จัดเตรียมไว้มาแบ่งออกเป็น 2 ส่วน โดย ส่วนที่ 1 คือ ข้อมูลที่ใช้สำหรับการเรียนรู้ (Training Data) จำนวน 69,990 รายการ ข้อมูลในส่วนนี้จะถูกนำไปป้อนให้กับแบบจำลองการเรียนรู้ (Training Model) เพื่อให้เครื่องคอมพิวเตอร์

ได้เรียนรู้และจดจำรูปแบบ ลักษณะ และความสัมพันธ์ของข้อมูล สำหรับนำไปใช้ในการจำแนกประเภท และนำไปสร้างเป็นแบบจำลองต้นไม้ตัดสินใน (Decision Tree Model) ส่วนที่ 2 คือ ข้อมูลที่ใช้สำหรับการทดสอบ (Testing Data) จำนวน 29,996 รายการ ข้อมูลในส่วนนี้จะถูกนำไปใช้ในการประเมินผลการเรียนรู้ของแบบจำลองการเรียนรู้ (Training Model) โดยการนำชุดคำสั่ง (Library) train_test_split ที่บรรจุอยู่ใน Sklearn (Pedregosa, 2011, pp.2825-2830) มาใช้ในการดำเนินการแบ่งข้อมูล โดยมีอัตราส่วนระหว่างข้อมูลที่ใช้ในการเรียนรู้ และข้อมูลที่ใช้ในการทดสอบ เป็น 70:30

4. การสร้างแบบจำลองเพื่อจำแนกประเภทโดยใช้กระบวนการต้นไม้ตัดสินใจแบบจำแนกประเภทและแบบถดถอย (Classification and Regression Tree: CART)

ในขั้นตอนนี้ผู้ดำเนินการวิจัยได้นำกระบวนการสร้างต้นไม้ตัดสินใจแบบจำแนกประเภทและแบบถดถอย มาใช้ในการวิเคราะห์ซึ่งสามารถประมวลผลได้จากข้อมูลที่อยู่ในแต่ละแอตทริบิว (Attribute) ในลักษณะ If – Else แบบไบนารี โดยการแตกกิ่งของต้นไม้ที่สามารถแตกกิ่งได้เพียง 2 กิ่ง แบบเวียนบังเกิด (Recursive) จนกว่าจะเข้าเงื่อนไขใดเงื่อนไขหนึ่ง ได้แก่ 1) ข้อมูลที่พิจารณาอยู่มีลักษณะเหมือนกัน หรือถูกจำแนกให้อยู่ในประเภทเดียวกัน และ 2) ไม่เหลือข้อมูลให้นำมาพิจารณาต่อแล้ว (Leo Breiman, 1984) ซึ่งได้ผลลัพธ์เป็นข้อมูลที่ถูกจำแนกประเภทบรรจุใน Node สุดท้าย โดยประมวลผลข้อมูลรายชื่อโดเมนที่ได้จากการแบ่งข้อมูลสำหรับเรียนรู้ ร้อยละ 70 จำนวน 69,990 รายการ เพื่อสร้างแบบจำลองการจำแนกประเภท (Classification Model) และทำการวัดประสิทธิภาพของแบบจำลองด้วยข้อมูลสำหรับทดสอบ ร้อยละ 30 จำนวน 29,996 รายการ โดยการใช้ค่าดัชนีจีนี่ (Gini Index) ในการพิจารณาแบ่งข้อมูล ซึ่งสามารถคำนวณค่าดัชนีจีนี่ได้จากสูตร ดังนี้

$$Gini = 1 - \sum_{i=1}^{n} pi^2$$

เมื่อ $$\Delta Gini(A) = Gini(D) - Gini_A(D)$$

และ $$Gini_A(D) = \frac{|D1|}{|D|} Gini(D) + \frac{|D2|}{|D|} Gini(D_2)$$

5. การประเมินประสิทธิภาพของแบบจำลอง

การประเมินประสิทธิภาพของการจำแนกประเภทข้อมูล (Classification Evaluate) โดยการนำข้อมูลที่ได้จากการแบ่งข้อมูลที่ใช้สำหรับการทดสอบร้อยละ 30 จำนวน 29,996 รายการ ซึ่งข้อมูลชุดนี้คือข้อมูลที่แบบจำลองไม่เคยรู้จักมาก่อน โดยใช้เมทริกซ์ความสับสน (Confusion Matrix) ในการประเมินประสิทธิภาพในด้านต่าง ๆ ประกอบด้วยค่าความถูกต้อง (Accuracy), ค่าความแม่นยำ (Precision), ค่าการระลึกได้ (Recall) และค่าเฉลี่ยแบบ Harmonic Mean (F1 Score) ซึ่งสามารถคำนวณได้จากสูตร ดังนี้

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

ค่าความแม่นยำ (Precision) $Pricision = \frac{TP}{(TP+FP)}$

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

ค่าเฉลี่ยแบบ Harmonic Mean (F1 Score) $F1\ Score = 2\frac{(Precision\ x\ Recall)}{(Precision+Recall)}$

ผลการวิจัย

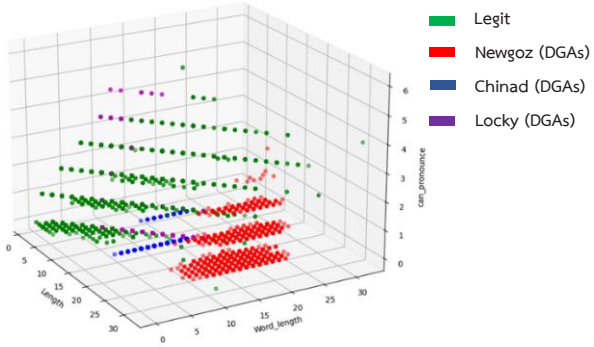
จากขั้นตอนในการดำเนินการวิเคราะห์ข้อมูลระหว่างชื่อโดเมนที่ถูกต้อง และชื่อโดเมนที่ถูกสร้างขึ้นมาจากอัลกอริทึมสำหรับสร้างโดเมน ประกอบด้วยผลการวิจัยที่สำคัญ ดังนี้

1. ค่าความบริสุทธิ์ (ΔGini)

ค่าความบริสุทธิ์ (ΔGini) ของข้อมูลสูงที่สุดจะถูกเลือกเป็นแอตทริบิวที่ใช้ในการแบ่งข้อมูล โดยสามารถคำนวณค่าความบริสุทธิ์ของข้อมูลได้จากสมการ (ΔGini) = Gini (D) – Gini $_A$ (A) เมื่อ D คือ ชุดข้อมูล และ A คือ แอตทริบิว ที่อยู่ในชุดข้อมูล (D) มีผลดังนี้

length	0.684238	elements	0.017456
word_length	0.164553	num_length	0.002676
can_pronounce	0.077410	elements_indict	0.002519
pronounce_elements	0.051017	in_dict	0.000130

เมื่อนำข้อมูลที่มีค่าความบริสุทธิ์ของข้อมูล (ΔGini) สูงที่สุด 3 อันดับแรก ได้แก่ length, word_length และ can_pronounce มาแสดงผลในรูปแบบของกราฟ 3 แนวแกน (3D Graph) จะสามารถมองเห็นรูปแบบการกระจายตัวของกลุ่มข้อมูลได้ดังภาพที่ 3



ภาพที่ 3 รูปแบบการกระจายตัวของกลุ่มข้อมูล โดยใช้แอตทริบิว (Attribute) ที่มีค่าความบริสุทธิ์ของข้อมูล (ΔGini) สูงที่สุด 3 อันดับแรก คือ length, word_length และ can_pronounce

2. ผลการประเมินประสิทธิภาพของแบบจำลองด้วยชุดข้อมูลการเรียนรู้ (Training Data)

จากข้อมูลรายชื่อโดเมนที่ใช้สำหรับการเรียนรู้ จำนวน 69,991 รายการ คิดเป็นร้อยละ 70 จากข้อมูลจำนวนทั้งสิ้น 99,987 รายการ สามารถนำมาสร้างเป็นแบบจำลองในการสร้างเงื่อนไขเพื่อจำแนกประเภทกลุ่มข้อมูลที่เป็นเป้าหมาย ได้แก่ Legit, Chinad, Locky และ Newgoz ซึ่งประกอบด้วยจำนวนชั้นความลึกทั้งหมด 16 ชั้น และมีเงื่อนไขที่ใช้ในการจำแนกประเภทของรายชื่อโดเมนจำนวน 508 เงื่อนไข สามารถแสดงผลการประเมินประสิทธิภาพของแบบจำลองในภาพที่ 4

report ข้อมูลเรียนรู้

	precision	recall	f1-score	support
chinad	1.00	0.99	1.00	12593
legit	0.98	0.97	0.97	34945
locky	0.87	0.93	0.90	9387
newgoz	1.00	1.00	1.00	13065
accuracy			0.97	69990
macro avg	0.96	0.97	0.97	69990
weighted avg	0.97	0.97	0.97	69990

Train set accuracy = 0.9719388484069152

ภาพที่ 4 แสดงผลการประเมินค่าประสิทธิภาพของแบบจำลองโดยใช้ชุดข้อมูลการเรียนรู้ (Training Data)

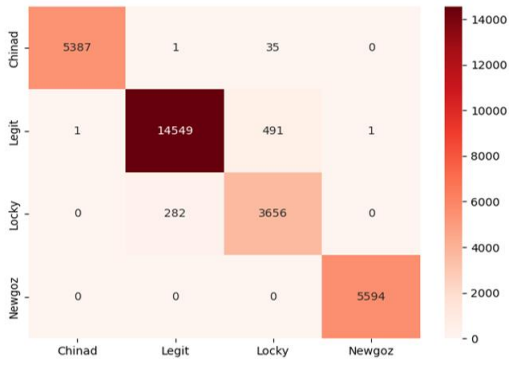
3. ผลการประเมินประสิทธิภาพของแบบจำลองด้วยชุดข้อมูลทดสอบ (Testing Data)

การประเมินประสิทธิภาพของแบบจำลองด้วยชุดข้อมูลสำหรับการทดสอบ จำนวน 29,997 รายการ คิดเป็นร้อยละ 30 จากข้อมูลจำนวนทั้งสิ้น 99,987 รายการ สามารถนำมาทดสอบความถูกต้องของแบบจำลองในการสร้างเงื่อนไขเพื่อจำแนกประเภทกลุ่มข้อมูลที่เป็นเป้าหมาย ได้แก่ Legit, Chinad, Locky และ Newgoz โดยใช้เมทริกซ์ความสับสน (Confusion Matrix) ในการประเมินประสิทธิภาพในด้านต่าง ๆ ประกอบด้วย ค่าความถูกต้อง (Accuracy), ค่าความแม่นยำ (Precision), ค่าการระลึกได้ (Recall) และค่าเฉลี่ยแบบ Harmonic Mean (F1 Score) โดยผลที่ได้จากการประเมินประสิทธิภาพของแบบจำลอง และจำนวนความถูกต้องในการจำแนกประเภทของแบบจำลองถูกแสดงในภาพที่ 5

report ข้อมูลทดสอบ

	precision	recall	f1-score	support
chinad	1.00	0.99	1.00	5423
legit	0.98	0.97	0.97	15042
locky	0.87	0.93	0.90	3938
newgoz	1.00	1.00	1.00	5594
accuracy			0.97	29997
macro avg	0.96	0.97	0.97	29997
weighted avg	0.97	0.97	0.97	29997

Test set accuracy = 0.972963963062973



ภาพที่ 5 แสดงผลการประเมินค่าประสิทธิภาพของแบบจำลองโดยใช้ชุดข้อมูลสำหรับการทดสอบ (Testing Data)

4. การปรับปรุงประสิทธิภาพของแบบจำลอง

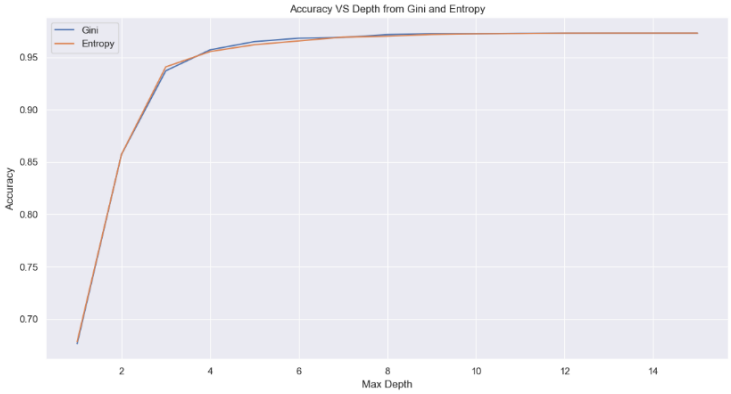
จากผลการวิจัยที่ได้นำเสนอนี้ ผู้ดำเนินการวิจัยได้ดำเนินการลดความซับซ้อนของแบบจำลองด้วยวิธีการตัดกิ่งในขณะเรียนรู้ (Pre Pruning) เนื่องจากแบบจำลองมีผลการทดสอบระหว่างข้อมูลที่นำมาใช้ในการเรียนรู้ และข้อมูลที่นำมาใช้ในการทดสอบ นั้น มีความสอดคล้องและใกล้เคียงกันมาก ซึ่งจากผลการวิจัยที่ได้รับจากชุดข้อมูลทั้งคู่มีความถูกต้อง (Accuracy) สอดคล้องกันสูงถึงร้อยละ 97 ซึ่งหมายความว่าแบบจำลองมีความเหมาะสม (Model Fit) ต่อการนำไปใช้งานแล้ว แต่เนื่องจากแบบจำลองมีชั้นความลึก (Max Depth) มากถึง 16 ชั้น จึงทำให้แบบจำลองมีความซับซ้อน และยากต่อการนำไปแปรผล อีกทั้งค่าความซับซ้อน (Alpha: α) ของแบบจำลองที่ประมวลผลได้โดยใช้วิธีการตัดกิ่งแบบค่าความซับซ้อน (Cost Complexity Pruning) มีค่าน้อยกว่าศูนย์ ($\alpha < 0$) (F. Esposito, Malerba, Semeraro, & Kay, 1997, pp.476-491) ส่งผลให้ไม่สามารถดำเนินการลดความซับซ้อนของแบบจำลองด้วยวิธีการตัดกิ่งแบบค่าความซับซ้อนได้ (Floriana Esposito, Malerba, Semeraro, & Tamma, 1999, pp.277-299) ดังแสดงผลการหาค่าความซับซ้อน (Alpha: α) ของแบบจำลองและผลความถูกต้องของแบบจำลองที่สอดคล้องกันระหว่างข้อมูลที่นำมาใช้ในการเรียนรู้ และข้อมูลที่นำมาใช้ในการทดสอบ ในภาพที่ 6

```
ตรวจสอบค่าความถูกต้องจากข้อมูลเรียนรู้
col_0          chinad  legit  locky  newgoz
subclass_type
chinad          12510      0     83       0
legit               7  33746   1192       0
locky               0    682   8705       0
newgoz              0      0      0   13065
Train set accuracy =0.9719388484069152
Test set accuracy =0.972963963062973

{'ccp_alphas': array([ 0.00000000e+00, -2.71050543e-20,  3.18323993e-09,  3.26849641e-09,
        8.14550306e-09,  9.29005167e-09,  9.39133243e-09,  1.25019062e-08,
        1.58650009e-08,  1.79758232e-08,  1.84403494e-08,  2.06048630e-08,
        3.15533590e-08,  3.35524588e-08,  4.75202508e-08,  4.89271783e-08,
```

ภาพที่ 6 แสดงรายการค่าความซับซ้อน (Cost Complexity Pruning) มีค่าน้อยกว่าศูนย์ ($\alpha < 0$) และผลความถูกต้อง (Accuracy) ของแบบจำลองที่สอดคล้องกันระหว่างข้อมูลที่นำมาใช้ในการเรียนรู้ และข้อมูลที่นำมาใช้ในการทดสอบ

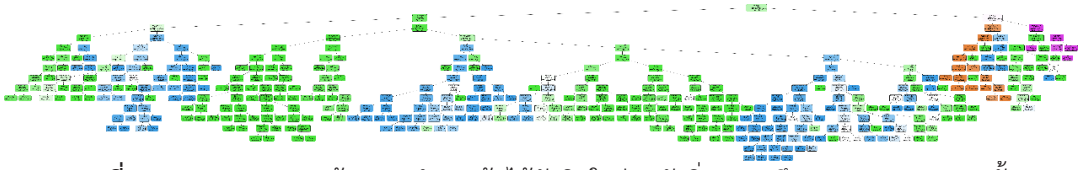
การตัดกิ่งในขณะเรียนรู้ดำเนินการโดยการเปรียบเทียบค่าความถูกต้องที่ได้จากเกณฑ์ที่ใช้ในการเลือกแอตทริบิวสำหรับสร้างแบบจำลองระหว่าง Gini และ Entropy เพื่อหาความสัมพันธ์ระหว่างค่าดังกล่าวกับชั้นความลึก (Max Depth) ของแบบจำลอง (Breslow & Aha, 1997, pp.1-40) จากการนำค่าดังกล่าวมาพล็อตเป็นกราฟ พบว่าค่าความถูกต้อง (Accuracy) ที่ได้จาก Gini (เส้นสีน้ำเงิน) และ Entropy (เส้นสีส้ม) เป็นเกณฑ์ที่ใช้ในการเลือกแอตทริบิวมีความสอดคล้องและใกล้เคียงกันมาก โดยที่ชั้นความลึกเท่ากับ 9 จะได้ค่าความถูกต้องสูงที่สุดอยู่ที่ร้อยละ 97 โดยประมาณ และจะคงค่านี้ไปเรื่อย ๆ จนถึงชั้นที่ 16 ดังแสดงในภาพที่ 7



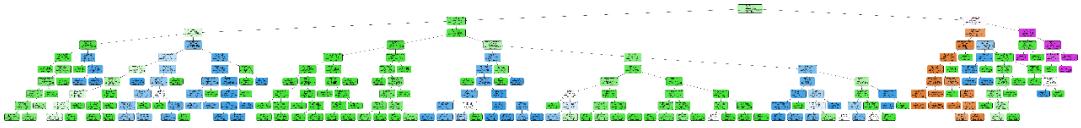
ภาพที่ 7 แสดงค่าความถูกต้อง (Accuracy) ที่ได้จากเกณฑ์ที่ใช้ในการเลือกแอตทริบิว (Attribute) ระหว่าง Gini และ Entropy เพื่อใช้ในการหาค่าชั้นความลึก (Max Depth) ที่เหมาะสม

ดังนั้น ผู้ดำเนินการวิจัยจึงดำเนินการปรับค่าชั้นความลึกของแบบจำลองต้นไม้ตัดสินใจ (Decision Tree) โดยกำหนดค่าชั้นความลึกเท่ากับ 9 และทำการประเมินประสิทธิ์ภาพของแบบจำลองด้วยชุดข้อมูลสำหรับทดสอบ โดยใช้เมทริกซ์ความสับสน (Confusion Matrix) ในการประเมินประสิทธิภาพในด้านต่าง ๆ พบว่ายังสามารถคงค่าความถูกต้องอยู่ที่ร้อยละ 97 โดยประมาณ

จากผลการวิจัยพบว่าเมื่อนำข้อมูลที่ใช้สำหรับการเรียนรู้มาสร้างเป็นแบบจำลอง โดยปรับค่าความลึก (Max Depth) เท่ากับ 9 จะเห็นได้ว่าค่าความซับซ้อนของแบบจำลองลดลงอย่างมีนัยสำคัญ และมีเงื่อนไขที่ใช้สำหรับการจำแนกประเภทของรายชื่อโดเมน จำนวน 340 เงื่อนไข ลดลงจากเดิม จำนวน 168 เงื่อนไข โดยที่ค่าความถูกต้อง (Accuracy) ของแบบจำลองที่ได้จาก Gini และ Entropy มีความสอดคล้องใกล้เคียงกันมาก และมีค่าความถูกต้อง สูงที่สุดอยู่ที่ร้อยละ 97 (โดยประมาณ) และจะคงค่านี้ไปเรื่อย ๆ จนถึงชั้นที่ 16 (มีค่าเพิ่มขึ้นเพียงเล็กน้อยในหลักทศนิยม) ซึ่งสามารถเปรียบเทียบลักษณะความซับซ้อนของแบบจำลองระหว่างก่อนตัดกิ่ง และหลังตัดกิ่งได้ดังภาพที่ 9



ภาพที่ 9 (ก) แสดงผลการสร้างแบบจำลองต้นไม้ตัดสินใจก่อนตัดกิ่ง ความลึก (Max Depth) 16 ชั้น



ภาพที่ 9 (ข) แสดงผลการสร้างแบบจำลองต้นไม้ตัดสินใจหลังตัดกิ่ง ความลึก (Max Depth) 9 ชั้น

ภาพที่ 9 แสดงผลการสร้างแบบจำลองต้นไม้ตัดสินใจแบบจำแนกประเภทและแบบถดถอย (CART) เปรียบเทียบระหว่างก่อน (ก) และหลังตัดกิ่ง (ข)

อภิปรายผล

1. ผลการเปรียบเทียบความถูกต้อง (Accuracy) ในการจำแนกประเภทของแบบจำลอง

แบบจำลองที่ถูกสร้างขึ้นโดยชุดข้อมูลสำหรับการเรียนรู้ เมื่อนำมาทดสอบด้วยชุดข้อมูลสำหรับการทดสอบจำนวน 29,997 รายการ โดยการปรับชั้นความลึกของแบบจำลองให้เหลือเพียง 9 ชั้น จากเดิม 16 ชั้น แบบจำลองสามารถจำแนกประเภทข้อมูลได้ถูกต้อง จำนวน 29,175 รายการ คิดเป็นร้อยละ 97.25 และจำแนกประเภทข้อมูลผิดจำนวน 822 รายการ หรือคิดเป็นร้อยละ 2.74 ดังแสดงในตารางที่ 1

ตารางที่ 1 เปรียบเทียบความถูกต้อง ในการจำแนกประเภทระหว่างก่อนและหลังการปรับชั้นความลึกของแบบจำลองทั้งแบบจำแนกประเภท และไม่จำแนกประเภทอัลกอริทึมสำหรับสร้างโดเมน

ผลการทำนาย / รูปแบบข้อมูล	จำนวนข้อมูลที่ใช้ในการทดสอบทั้งหมด	ทำนายถูก		ทำนายผิด		ความถูกต้อง (Accuracy)	
		ความลึก 16 ชั้น	ความลึก 9 ชั้น	ความลึก 16 ชั้น	ความลึก 9 ชั้น	ความลึก 16 ชั้น	ความลึก 9 ชั้น
1.1 แบบจำแนกประเภทของอัลกอริทึมสำหรับสร้างโดเมน							
Legitimate Domain Name	15,042	14,549	14,535	493	507	96.72	96.62
Chinad (DGA)	5,423	5,387	5,387	36	36	99.33	99.33
Locky (DGA)	3,938	3,656	3,659	282	279	92.83	92.91
Newgoz (DGA)	5,594	5,594	5,594	0.00	0.00	100.00	100.00
รวมทั้งสิ้น	**29,997**	**29,186**	**29,175**	**811**	**822**	**97.29**	**97.25**
1.2 แบบไม่จำแนกประเภทของอัลกอริทึมสำหรับสร้างโดเมน							
Legitimate Domain Name	15,042	14,549	14,535	493	507	96.72	96.62
DGA	14,955	14,637	14,640	318	315	97.38	97.41
รวมทั้งสิ้น	**29,997**	**29,186**	**29,175**	**811**	**822**	**97.29**	**97.25**

2. ผลการวัดประสิทธิภาพของแบบจำลองเปรียบเทียบระหว่างก่อนและหลังการปรับชั้นความลึกของแบบจำลอง

การวัดประสิทธิภาพของแบบจำลองด้วยชุดข้อมูลสำหรับการทดสอบ โดยใช้เมทริกซ์ความสับสนในการประเมินประสิทธิภาพในด้านต่าง ๆ เปรียบเทียบระหว่างก่อนและหลังการปรับชั้นความลึกของแบบจำลองเพื่อลดความซับซ้อน โดยการแยกประเภทข้อมูลสามารถแสดงผลได้ตามตารางที่ 2

ตารางที่ 2 แสดงผลการวัดประสิทธิภาพโดยใช้เมทริกซ์ความสับสน เปรียบเทียบระหว่างก่อนและหลังการปรับชั้นความลึกของแบบจำลองทั้งแบบจำแนกประเภท และไม่จำแนกประเภทอัลกอริทึมสำหรับสร้างโดเมน

ผลการวัดประสิทธิภาพ / รูปแบบข้อมูล	จำนวนข้อมูลที่ใช้ในการทดสอบทั้งหมด	ค่าความแม่นยำ (Precision)		ค่าการระลึกได้ (Recall)		ค่าเฉลี่ยแบบ Harmonic Mean (F1 Score)	
		ความลึก 16 ชั้น	ความลึก 9 ชั้น	ความลึก 16 ชั้น	ความลึก 9 ชั้น	ความลึก 16 ชั้น	ความลึก 9 ชั้น
2.1 แบบจำแนกประเภทของอัลกอริทึมสำหรับสร้างโดเมน							
Legitimate Domain Name	15,042	98.00	98.00	97.00	97.00	97.00	97.00
Chinad (DGA)	5,423	100.00	100.00	99.00	99.00	100.00	100.00
Locky (DGA)	3,938	87.00	87.00	93.00	93.00	90.00	90.00
Newgoz (DGA)	5,594	100.00	100.00	100.00	100.00	100.00	100.00
เฉลี่ย	**29,997**	**96.25**	**96.25**	**97.25**	**97.25**	**96.75**	**96.75**
2.2 แบบไม่จำแนกประเภทของอัลกอริทึมสำหรับสร้างโดเมน							
Legitimate Domain Name	15,042	98.00	98.00	97.00	97.00	97.00	97.00
DGA	14,955	95.66	95.66	97.33	97.33	96.66	96.66
เฉลี่ย	**29,997**	**96.83**	**96.83**	**97.16**	**97.16**	**96.83**	**96.83**

กิตติกรรมประกาศ

ผู้วิจัยขอขอบพระคุณคณะอาจารย์ที่ปรึกษาทุกท่านที่เสียสละเวลาอันมีค่า ทุ่มเทแรงกาย แรงใจ ให้คำชี้แนะ และคำปรึกษาทั้งในด้านวิชาการ และการดำเนินชีวิตทั้งทางตรงและทางอ้อม รวมถึงมหาวิทยาลัยรังสิตที่ให้การสนับสนุนด้านงบประมาณในการจัดทำการวิจัย

เอกสารอ้างอิง

Bader, J. (2015). **Domain Generation Algorithms (DGAs) of Malware reimplemented in Python.** [Online]. Available : https://github.com/baderj/domain_generation_algorithms [2021, October, 23].

Bird, S., Klein, E., & Loper, E. (2009). **Natural Language Processing with Python**. O'Reilly Media, Inc.

Breslow, L. A., & Aha, D. W. (1997). Simplifying decision trees: A survey. **The Knowledge Engineering Review, 12**(01), 1-40.

Chowdhury, S. A. (2019). Domain Generation Algorithm-Dga in Malware.

Esposito, F., Malerba, D., Semeraro, G., & Kay, J. (1997). A comparative analysis of methods for pruning decision trees. **IEEE Transactions on Pattern Analysis and Machine Intelligence, 19**(5), 476-491.

Esposito, F., Malerba, D., Semeraro, G., & Tamma, V. (1999). The effects of pruning methods on the predictive accuracy of induced decision trees. **Applied Stochastic Models in Business and Industry, 15**(4), 277-299.

G. P., A., R., G., S., K., & Gladston, A. (2020). A machine learning framework for domain generating algorithm based malware detection. **Security and Privacy, 3**(6), e127.

Hubbard, D. (2016). **Cisco Umbrella 1 Million**. [Online]. Available : https://umbrella.cisco.com/blog/cisco-umbrella-1-million [2021, April, 12].

Jenks, G. (2018). **Python Word Segmentation**. [Online]. Available : https://github.com/grantjenks/python-wordsegment.git [2021, May, 16].

Kurkowski, J. (2020). **tldextract**. [Online]. Available : https://github.com/john-kurkowski/tldextract.git [2021, June, 3].

Leo Breiman, J. H. F., Richard A. Olshen, Charles J. Stone. (1984). **Classification And Regression Trees s. Edition *(Ed.)***. [Online]. Available : https://doi.org/10.1201/9781315139470 [2021, May, 4].

Parrish, A. (2015, November). **pronouncingpy**. [Online]. Available : https://github.com/aparrish/pronouncingpy.git [2021, July, 15].

Pedregosa, F. a. V., G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research, 12**, 2825-2830.

The European Union Agency for Cybersecurity, E. (2020). **List of top 15 threats. ENISA Threat Landscape** [Online]. Available : https://www.enisa.europa.eu/publications/enisa-threat-landscape-2020-list-of-top-15-threats/view/++widget++form.widgets.fullReport/@@download/ETL2020+-+ENISA+List+ of+top+15+Threats+A4.pdf [2021, July, 9].