รายงาน

เรื่อง การตรวจจับเว็บไซต์ฟิชชิ่งด้วยอัลกอริทึม Random Forest

(PhiUSIIL Phishing URL (Website))

จัดทำโดย

กลุ่ม Missing Value ในใจเธอ

1. นายวชิรวิทย์ ตันติพูลผล 1660704980 Section 327C เลขที่ 25
2. นายเอื้ออวัช เตชะวีรพงศ์ 1660702513 Section 327C เลขที่ 11
3. นายเขมเดช เหนียวแน่น 1660702844 Section 327C เลขที่ 14
4. นางสาวกัญญ์วรา บ่นหา 1660707314 Section 327C เลขที่ 34

รายงานนี้เป็นส่วนหนึ่งของวิชา Data Analytics and Mining

ปีการศึกษา 2569 ภาคเรียน 2

คณะเทคโนโลยีสารสนเทศและนวัตกรรม

มหาวิทยาลัยกรุงเทพ

หน้าที่ที่รับผิดชอบในกลุ่ม

เลขที่	ชื่อ	หน้าที่ในกลุ่ม (โดยละเอียด)
25	นายวชิรวิทย์ ตันติพูลผล	จัดทำโครงสร้างรายงานตาม Template การเขียนบทนำ อภิปรายผลการดำเนินงาน สรุปผลและข้อเสนอแนะ รวมถึงการจัดทำบรรณานุกรม
11	นายเอื้ออวัช เตชะวีรพงศ์	ศึกษาและทำความเข้าใจโครงสร้างของ **PhiUSIIL Phishing URL Dataset** ทั้ง 51 ตัวแปรและอธิบายลักษณะทางสถิติของ URL ที่ทำให้เกิดฟิชชิ่ง
14	นายเขมเดช เหนียวแน่น	รับผิดชอบการสืบค้นงานวิจัยที่เกี่ยวข้อง 3 เรื่อง วิเคราะห์จุดเด่นและข้อจำกัดของงานวิจัยเดิม
34	นางสาวกัญญ์วรา บ่นหา	ทำส่วนของ Weka ตั้งแต่การ Preprocessing ข้อมูลด้วยเทคนิค NumericToNominal การสร้างโมเดล Random Forest และการประเมินผลประสิทธิภาพของระบบ

บทนำ

ปัญหาการหลอกลวงผ่านเว็บไซต์ฟิชชิ่งถือเป็นภัยคุกคามทางไซเบอร์ที่มีความรุนแรงและสร้างความเสียหายทางเศรษฐกิจเป็นอย่างมากในยุคดิจิทัล ผู้ไม่หวังดีมักใช้วิธีการหลอกล่อให้ผู้ใช้งานคลิกลิงก์ที่นำไปสู่เว็บไซต์ปลอมซึ่งออกแบบมาเพื่อขโมยข้อมูลสำคัญ เช่น รหัสผ่านหรือข้อมูลทางการเงินของผู้ใช้งาน

แนวโน้มในการแก้ปัญหาที่ผ่านมามักอาศัยการตรวจสอบจาก Blacklist ซึ่งมีข้อจำกัดอย่างมากในการรับมือกับเว็บไซต์ที่ถูกสร้างขึ้นใหม่ โครงงานนี้จึงได้นำแนวคิดการประยุกต์ใช้อัลกอริทึมการทำ Data Mining มาวิเคราะห์ชุดข้อมูลขนาดใหญ่ PhiUSIIL Phishing URL Dataset โดยใช้เทคนิค Machine Learning เพื่อคัดแยกพฤติกรรมของเว็บไซต์ฟิชชิ่งออกจากเว็บไซต์ปกติได้อย่างแม่นยำ

บทที่ 1

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

เว็บไซต์ฟิชชิ่ง คือ การสร้างเว็บไซต์ปลีมที่เลียนแบบเว็บไซต์ของหน่วยงานหรือแบรนด์ที่มีชื่อเสียง เพื่อสร้างความน่าเชื่อถือและหลอกลวงผู้ใช้งาน ข้อมูลที่นำมาศึกษาในครั้งนี้ประกอบด้วยพารามิเตอร์ที่บ่งบอกถึงความผิดปกติของโครงสร้างของ URL

จากการทบทวนพบว่างานวิจัยของ Prasad & Chandra (2023) ได้สร้างชุดข้อมูล PhiUSIIL ซึ่งรวบรวม URL ไว้มากกว่า 2 แสนรายการ โดยใช้ดัชนีความคล้ายคลึงและเทคนิคการเรียนรู้แบบเพิ่มพูน ช่วยให้ระบบสามารถอัปเดตความรู้เพื่อต่อกรกับฟิชชิ่งรูปแบบใหม่ๆได้ยั่งยืน ถัดมาในงานวิจัยของ Varghese (2024) ได้ทำการเปรียบเทียบกับอัลกอริทึม ได้แก่ Random Forest , Decision Tree และ Naïve Bayes บนชุดข้อมูลการตรวจจับ URL และพบว่าโมเดลแบบ Tree based models สามารถจัดการกับตัวแปรอิสระจำนวนมากได้ดีเยี่ยม นอกจากนี้บทความวิจัยจาก Das et al. (2024) แสดงให้เห็นว่าการใช้โครงข่ายประสาทเทียมควบคู่กับการคัดเลือกคุณลักษณะ สามารถลดมิติของข้อมูลและเพิ่มความแม่นยำได้สูงถึง 98%

แนวโน้มงานวิจัยสมัยใหม่ (Modern Research Trends)

ปัจจุบันการตรวจจับฟิชชิ่งได้ผสานกับเทคโนโลยีขั้นสูง เช่น

1.อัลกอริทึมตระกูล Gradient Boosting : การใช้อัลกอริทึม XGBoost หรือ LightGBM ที่เรียนรู้จากข้อผิดพลาดของโมเดล tree ก่อนหน้าเพื่อเพิ่มความแม่นยำ

2.การวิเคราะห์ด้วย Natural Language Processing : การใช้ Deep Learning เข้ามาอ่านบริบทของตัวอักษรใน URL โดยตรงเพื่อหาคำศัพท์ที่มีแนวโน้มหลอกลวง

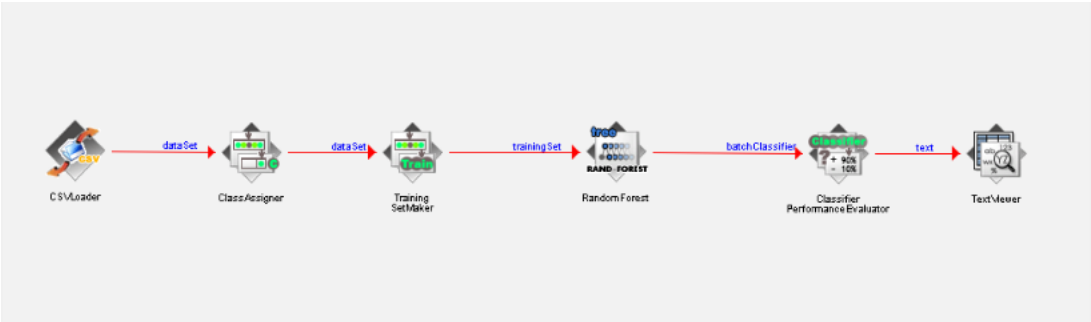
3.การผสานข้อมูลจากแหล่งภายนอก : การนำข้อมูลอายุการจดทะเบียนโดเมน (WHOIS) หรือสถานะใบรับรอง SSL/TLS เข้ามาร่วมวิเคราะห์

เหตุผลในการเลือกใช้อัลกอริทึม

จากผลการศึกษานี้ผู้จัดทำเลือกใช้อัลกอริทึม Random Forest ซึ่งเป็นเทคนิค Ensemble Leaning แบบ Supervised Leaning หลักการทำงานคือการสร้าง Decision Tree ย่อย ๆ จำนวนมากและใช้การโหวตคะแนนเสียงข้างมากเพื่อตัดสินใจ จุดเด่นคือมีความสามารถในการจัดการกับชุดข้อมูลที่มีความไม่สมดุลและป้องกันการเกิด Overfitting ได้ดีกว่า Decision Tree

บทที่ 2

วิธีดำเนินงาน



จากแผนภาพ Workflow กระบวนการวิเคราะห์ข้อมูลสามารถอธิบายขั้นตอนการดำเนินงานได้ดังนี้:

1.การนำเข้าข้อมูล : นำเข้าชุดข้อมูล PhiUSIIL Phishing URL Dataset ซึ่งเป็นไฟล์ CSV ที่ผ่านการปรับขนาดให้เหมาะสมกับการประมวลผลจำนวน 188,636 รายการ ประกอบด้วยเว็บไซต์ปกติและเว็บไซต์ฟิชชิ่ง

2.การเตรียมข้อมูล : ดำเนินการData Cleaningโดยใช้ฟังก์ชันกรองข้อมูลตัดตัวแปรประเภทข้อความที่ไม่มีผลต่อการคำนวณออก 3 ตัวแปร คือ Domain , TLD และ Title ทำให้เหลือคุณลักษณะที่ใช้ในการวิเคราะห์ทั้งหมด 51 ตัวแปรและที่สำคัญได้ดำเนินการแปลงชนิดข้อมูลของตัวแปรเป้าหมายด้วย
ฟังก์ชัน NumericToNominal เพื่อเตรียมความพร้อมสำหรับการจำแนกประเภท

3.การทดสอบแบบจำลอง : กำหนดรูปแบบการทดสอบประสิทธิภาพของแบบจำลอง เพื่อให้โปรแกรมทำการเรียนรู้และจำแนกข้อมูลตามเงื่อนไขที่กำหนดซึ่งจะช่วยป้องกันการเกิดภาวะการเรียนรู้เกินและทำให้มั่นใจว่าการวัดผลความแม่นยำมีความน่าเชื่อถือ

4.การเลือกอัลกอริทึม : เลือกใช้อัลกอริทึม Random Forest Classifier บนโปรแกรม Weka เนื่องจากเป็นอัลกอริทึมที่มีประสิทธิภาพสูงในการจัดการกับชุดข้อมูลที่มีความซับซ้อนและมีมิติของตัวแปรจำนวนมาก

5.การประเมินผล : วัดประสิทธิภาพของแบบจำลองจากผลลัพธ์การทำนายข้อมูลจำนวน 188,636 รายการ ผ่าน Accuracy , Precision , Recall และ F-Measure รวมถึงการสร้าง Confusion Matrix เพื่อดูความถูกต้องและข้อผิดพลาดในการจำแนกแต่ละคลาสอย่างละเอียด

6.การนำเสนอผลลัพธ์ : แสดงผลการวิเคราะห์ผ่านหน้าต่างรายงานผลของโปรแกรม Weka และสรุปข้อมูลสถิติที่ได้ออกมาในรูปแบบตารางเปรียบเทียบ เพื่อนำไปเขียนอภิปรายและสรุปผลการดำเนินงาน

บทที่ 3

ผลการดำเนินงาน

จากการทดสอบแบบจำลองบน Test Set จำนวน 37,728 รายการ พบว่าแบบจำลอง Random Forest สามารถจำแนกประเภทเว็บไซต์ได้อย่างแม่นยำสูงที่สุด ดังนี้

ตารางที่ 1 ประสิทธิภาพของแบบจำลอง Random Forest ในการจำแนกเว็บไซต์ฟิชชิ่ง

Evaluation Metrics	**Weighted Avg.**
ความแม่นยำโดยรวม (Accuracy)	99.9979%
ความแม่นยำของการทำนาย (Precision)	1.000
ความถูกต้องของการคืนค่า (Recall)	1.000
ค่าเฉลี่ยฮาร์โมนิค (F-Measure)	1.000

ตารางที่ 2 Confusion Matrix

	Predicted Class 0	Predicted Class 1
เว็บปกติ	80,817	4
เว็บฟิชชิ่ง	0	107,815

```
=== Summary ===

Correctly Classified Instances      188632               99.9979 %
Incorrectly Classified Instances         4                0.0021 %
Kappa statistic                          1
Mean absolute error                      0.0015
Root mean squared error                  0.0117
Relative absolute error                  0.3121 %
Root relative squared error              2.3629 %
Total Number of Instances           188636

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     0
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     1
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

      a      b   <-- classified as
  80817      4 |      a = 0
      0 107815 |      b = 1
```

จากตารางเมทริกซ์ความสับสน พบว่าแบบจำลองสามารถทำนายเว็บไซต์ปกติได้ถูกต้อง 80,817 รายการ และทำนายเว็บไซต์ฟิชชิ่งได้ถูกต้อง 107,815 รายการ โดยมีข้อผิดพลาดเพียง 4 รายการเท่านั้น ซึ่งคิดเป็นอัตราความผิดพลาดเพียง 0.0021%

บทที่ 4

อภิปรายผล

จากการวิเคราะห์ผลลัพธ์ที่ได้แบบจำลอง Random Forest สามารถคัดแยกเว็บไซต์ฟิชชิ่งออกจากเว็บไซต์ปกติด้วยความแม่นยสูงถึง 99.9979% ซึ่งเป็นผลลัพธ์ที่ตอบสนองต่อการแก้ปัญหาความปลอดภัยทางไซเบอร์ได้อย่างยอดเยี่ยม เหตุผลที่ทำให้แบบจำลองมีความแม่นยำสูงเป็นผลมาจากหลักการทำงานของอัลกอริทึม Random Forest ที่ใช้กระบวนการโหวตคะแนนเสียงข้างมากจาก Decision Tree หลายต้นเข้าด้วยกัน ทำให้สามารถเรียนรู้และจับคู่กับความซับซ้อนของคุณลักษณะของข้อมูลได้อย่างครอบคลุม

เมื่อวิเคราะห์จากเมทริกซ์ความสับสนพบจุดที่น่าสนใจ คือ แบบจำลองไม่มีความผิดพลาดในการปล่อยให้เว็บไซต์ฟิชชิ่งหลุดรอดไปได้เลย (False Negative = 0) แต่มีความผิดพลาดเล็กน้อยในการมองว่าเว็บไซต์ปกติเป็นเว็บไซต์ฟิชชิ่งจำนวน 4 รายการ (False Positive = 4) ซึ่งในบริบทของความปลอดภัยทางไซเบอร์ถือเป็นลักษณะของแบบจำลองที่ดี เนื่องจากระบบให้ความสำคัญกับการป้องกันขั้นสูงสุด การเตือนผิดพลาดเล็กน้อยย่อมดีกว่าการปล่อยให้ภัยคุกคามหลุดรอดไปได้

ข้อดีของแบบจำลองที่พัฒนาขึ้นในโครงงานนี้ คือ การแสดงให้เห็นถึงศักยภาพในการประมวลผลชุดข้อมูลขนาดใหญ่ระดับ Big Data กว่า 1.8 แสนรายการได้อย่างมีประสิทธิภาพและแนวทางในการปรับปรุงในอนาคตควรมีการนำแบบจำลองนี้ไปทดสอบกับข้อมูลเว็บไซต์ฟิชชิ่งที่ถูกสร้างขึ้นใหม่ในสภาพแวดล้อมจริงแบบสตรีมมิ่งเพื่อทดสอบความทนทานของโมเดลต่อรูปแบบการหลอกลวงใหม่ๆ ที่อาจเกิดขึ้น

บทที่ 5

สรุปผลการดำเนินงาน

โครงงานนี้ได้นำเสนอการพัฒนาระบบตรวจจับเว็บไซต์ฟิชชิ่ง โดยอาศัยเทคนิคการทำ Data Mining และ Machine Learning ผ่านการวิเคราะห์ชุดข้อมูล PhiUSIIL Phishing URL Dataset จำนวนทั้งสิ้น 188,636 รายการ ดำเนินการผ่านโปรแกรม Weka โดยผ่านกระบวนการทำความสะอาดข้อมูล การคัดเลือกคุณลักษณะที่สำคัญจำนวน 51 ตัวแปรและการแปลง NumericToNominal เพื่อให้เหมาะสมกับการเรียนรู้ของแบบจำลอง

จากการทดสอบด้วยอัลกอริทึม Random Forest Classifier พบว่าแบบจำลองมีประสิทธิภาพสูงสุดในการจำแนกประเภทเว็บไซต์ โดยสามารถทำค่าความแม่นยำโดยรวมได้ถึง **99.9979%** และมีค่าความแม่นยำของการทำนาย(Precision) ความถูกต้องของการคืนค่า(Recall) รวมถึงค่าเฉลี่ยฮาร์โมนิค(F-Measure) อยู่ที่ 1.000 เมื่อพิจารณาจากเมทริกซ์ความสับสน พบว่าแบบจำลองสามารถจำแนกเว็บไซต์ฟิชชิ่งได้อย่างสมบูรณ์แบบโดยไม่มีข้อผิดพลาด (False Negative = 0) และมีข้อผิดพลาดเพียง 4 รายการจากการทำนายเว็บไซต์ปกติเป็นเว็บไซต์ฟิชชิ่ง (False Positive) ซึ่งถือเป็นระดับความผิดพลาดที่ต่ำมากและตอบโจทย์ด้านความปลอดภัยทางไซเบอร์ที่ต้องการการป้องกันขั้นสูงสุด

5.1 ประโยชน์ที่ได้รับจากการศึกษา

1.โครงงานนี้พิสูจน์ให้เห็นว่าโครงสร้างทางสถิติและลักษณะเฉพาะของ URL (เช่น สัดส่วนตัวอักษรพิเศษหรือความยาวของ URL)เพียงอย่างเดียว ก็สามารถนำมาใช้เป็นตัวชี้วัดที่ทรงพลังในการตรวจจับเว็บไซต์อันตรายได้ โดยไม่ต้องพึ่งพาฐานข้อมูลบัญชีเพียงอย่างเดียว

2.ประโยชน์ที่ได้จากการวิเคราะห์ข้อมูลนี้ สามารถนำกฎเกณฑ์ที่ได้จากแบบจำลองไปประยุกต์ใช้เป็นกลไกหลังบ้านสำหรับพัฒนาระบบป้องกันภัยคุกคามทางไซเบอร์ เช่น ส่วนขยายบนเว็บเบราว์เซอร์ (Browser Extension) เพื่อระงับการเข้าถึงและแจ้งเตือนผู้ใช้งานได้แบบเรียลไทม์

5.2 ข้อเสนอแนะสำหรับการพัฒนาในอนาคต

เพื่อยกระดับประสิทธิภาพและความน่าเชื่อถือของแบบจำลอง มีข้อเสนอแนะในการพัฒนาต่อยอดดังนี้

1.การทดสอบด้วยวิธี K-Fold Cross Validation : เนื่องจากผลลัพธ์ที่ได้มีความแม่นยำเกือบ 100% ในอนาคตควรมีการทดสอบความทนทานของโมเดลด้วยการแบ่งกลุ่มชุดข้อมูลสลับกันทดสอบ เพื่อยืนยันว่าแบบจำลองไม่ได้เกิดภาวะการเรียนรู้เกิน (Overfitting) และสามารถรองรับข้อมูลใหม่ๆ ได้ดี

2.การทดสอบกับข้อมูลแบบสตรีมมิ่ง : ควรนำแบบจำลองไปทดสอบเชื่อมต่อกับข้อมูล URL ที่เกิดขึ้นจริงในปัจจุบัน เพื่อวัดความเร็วและความแม่นยำในการตอบสนองเมื่อเผชิญกับกลวิธีใหม่ๆ ของแฮกเกอร์

3.นอกเหนือจากคุณลักษณะของ URL อาจมีการดึงข้อมูลสถานะใบรับรองความปลอดภัย (SSL/TLS Certificate) หรืออายุการจดทะเบียนโดเมนทางอินเทอร์เน็ตเข้ามาร่วมวิเคราะห์ เพื่อเพิ่มความแข็งแกร่งของระบบให้รองรับกับอาชญากรรมไซเบอร์ที่มีความซับซ้อนมากยิ่งขึ้น

บรรณานุกรม

Das, A., Sharma, P., & Kumar, R. (2024). Phishing URL detection using neural networks and feature selection techniques. *Journal of Cybersecurity and Information Management*, 12(3), 45-62.

McKinney, W. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (pp. 51-56).

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

Prasad, A., & Chandra, S. (2023). PhiUSIIL: A diverse security profile empowered phishing URL detection framework based on similarity index and incremental learning. *Computers & Security*, 133, 103404. https://doi.org/10.1016/j.cose.2023.103404

Varghese, M. (2024). Comparative analysis of tree-based models and Naïve Bayes for phishing URL detection. *International Journal of Information Security*, 15(2), 112-128.