

การตรวจสอบความผิดปกติของระบบบัตรเครดิตในเครือข่ายอีดีซี โดยใช้วิธีแบบป่าสุ่ม

Anomaly Credit Card System Detection in EDC Network Using Random Forest Algorithm

ณภัทร ไช่มุกด์ และ นิติเศษฐ์ หมวดทองอ่อน*
Napat Kaimuk & Nithizethe Mhuadthongon*

แขนงวิชาเทคโนโลยีดิจิทัล สาขาวิชาวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยสุโขทัยธรรมาธิราช
Digital Technology, School of Science and Technology, Sukhothai Thammathirath Open University

Submitted 10/09/2025 ; Revised 17/10/2025 ; Accepted 05/11/2025

บทคัดย่อ

ธุรกิจบัตรเครดิตในประเทศไทยมีแนวโน้มเติบโตอย่างต่อเนื่อง โดยคาดว่าจะยอดการใช้ผ่านบัตรเครดิตจะขยายตัวเฉลี่ยร้อยละ 8-9 ต่อปีในช่วงปี 2567-2568 อย่างไรก็ตาม ระบบเครือข่ายสำหรับการทำธุรกรรมยังคงประสบปัญหาการขัดข้องบ่อยครั้ง ส่งผลกระทบต่อผู้ใช้บริการทั้งด้านความสะดวก ความเชื่อมั่น และประสิทธิภาพในการทำงาน เครื่องมือที่ใช้อยู่ในปัจจุบันยังไม่สามารถตรวจสอบและตอบสนองได้อย่างทันทั่วถึง ทำให้ปัญหาการขัดข้องของระบบยังคงเกิดขึ้นอย่างต่อเนื่อง งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองการเรียนรู้ของเครื่องที่เหมาะสมสำหรับการจำแนกเหตุการณ์ผิดปกติในระบบบัตรเครดิตเครือข่ายอีดีซี โดยใช้ข้อมูลบันทึกเหตุการณ์ในอดีตร่วมกับข้อมูลแฟ้มล็อกจากอุปกรณ์อีดีซี รวมทั้งสิ้น 5,011 ชั่วโมง อัลกอริทึมที่นำมาพิจารณา ได้แก่ ต้นไม้ตัดสินใจ ป่าสุ่ม และเพื่อนบ้านใกล้เคียงที่สุด ผลการประเมินประสิทธิภาพแสดงให้เห็นว่า แบบจำลองที่ใช้อัลกอริทึมป่าสุ่มสามารถให้ค่าความถูกต้อง 99.96% ความแม่นยำ 99.90% ความครอบคลุม 99.90% และค่าประสิทธิภาพโดยรวม 99.90% สรุปได้ว่าอัลกอริทึมป่าสุ่มเป็นวิธีการที่เหมาะสมที่สุดในการพัฒนาเป็นต้นแบบสำหรับเครื่องมือตรวจสอบความผิดปกติของระบบบัตรเครดิตในเครือข่ายอีดีซี เพื่อยกระดับเสถียรภาพและความน่าเชื่อถือของระบบธุรกรรมทางการเงิน

คำสำคัญ: การเรียนรู้ของเครื่อง การตรวจสอบความผิดปกติ แฟ้มล็อก เครือข่ายอีดีซี

*ผู้ประสานงานหลัก (Corresponding Author)

E-mail: Nithizethe.Mhu@stou.ac.th

Abstract

The credit card business in Thailand continues to show steady growth, with spending projected to increase by an average of 8–9 percent annually during 2024–2025. However, the transaction network still faces frequent disruptions, which directly affect users in terms of convenience, trust, and operational efficiency. Existing monitoring tools cannot provide timely responses, resulting in recurring system malfunctions. This study aims to develop an appropriate machine learning model for classifying abnormal events in the credit card Electronic Data Capture (EDC) network. To achieve this, historical event records were combined with log files from EDC devices, totaling 5,011 hours of operational data. The algorithms considered included Decision Tree, Random Forest, and K-Nearest Neighbors. The performance evaluation revealed that the Random Forest algorithm achieved the best results, with accuracy, precision, recall, and overall efficiency of 99.96%, 99.90%, 99.90%, and 99.90%, respectively. The results indicate that the Random Forest algorithm is the most suitable method for developing a prototype tool to detect anomalies in credit card systems within the EDC network, thereby enhancing the stability and reliability of financial transactions.

Keywords: machine learning, anomaly detection, log file, EDC network

บทนำ

ในปัจจุบันนวัตกรรมทางการเงินของโลกมีการเปลี่ยนแปลง และพัฒนาอย่างต่อเนื่อง โดยได้รับแรงขับเคลื่อนจากเทคโนโลยีดิจิทัลแบบก้าวกระโดด ส่งผลให้เกิดบริการทางการเงินดิจิทัล (digital financial service) รูปแบบใหม่ ๆ ขึ้น เช่น กระเป๋าเงินดิจิทัล (e-wallet) สกุลเงินดิจิทัล (digital currencies) และการชำระเงินทางอิเล็กทรอนิกส์ (e-payment) [1] โดยเฉพาะอย่างยิ่ง การชำระเงินทางอิเล็กทรอนิกส์ ซึ่งถือเป็นระบบที่มีความสำคัญต่อการดำรงชีวิตในยุคดิจิทัล เนื่องจากช่วยให้ธุรกรรมทางการเงินและกิจกรรมทางเศรษฐกิจดำเนินไปได้อย่างสะดวก รวดเร็ว และมีประสิทธิภาพ อีกทั้งยังมีบทบาทสำคัญในการยกระดับคุณภาพชีวิตของประชาชน และเสริมสร้างเสถียรภาพให้กับระบบบริหารจัดการการเงินการคลังของประเทศ นอกจากนี้ การพัฒนาโครงสร้างพื้นฐานของระบบการชำระเงินทางอิเล็กทรอนิกส์ ภายใต้แผนยุทธศาสตร์การพัฒนาโครงสร้างพื้นฐานระบบการชำระเงินแห่งชาติ (national e-payment master plan) ยังเป็นปัจจัยสำคัญที่ช่วยเพิ่มขีดความสามารถในการแข่งขันของประเทศ ทั้งในด้านความสามารถในการแข่งขัน (competitiveness) ความยาก-ง่ายในการประกอบธุรกิจ (ease of doing business) และดัชนีการพัฒนามนุษย์ (human development index) [2]

จำนวนการใช้บัตรเครดิตได้รับความนิยมเพิ่มขึ้นอย่างต่อเนื่อง เนื่องจากมีความสะดวกสบาย สามารถสะสมคะแนนและสิทธิพิเศษ ช่วยบริหารกระแสเงินสด และยังมาพร้อมกับโปรโมชั่นหรือข้อเสนอพิเศษต่าง ๆ [3] ข้อมูลจากธนาคารแห่งประเทศไทยระบุว่า ในปี 2558 มีผู้ถือบัตรเครดิตจำนวน 21.8 ล้านใบ [4] ซึ่งสะท้อนให้เห็นถึงการเติบโตของการใช้บัตรเครดิตอย่างต่อเนื่อง อย่างไรก็ตาม การใช้งานที่เพิ่มสูงขึ้น โดยเฉพาะในช่วงเทศกาลสำคัญต่าง ๆ ที่มีธุรกรรมจำนวนมาก มักทำให้ระบบขาดเสถียรภาพ และเมื่อเกิดความผิดปกติ ระบบจะหยุดชะงักจนไม่สามารถทำงานได้อย่างต่อเนื่อง ปัญหานี้จึงเป็นเหตุผลสำคัญที่ต้องให้ความสำคัญกับการบริหารจัดการสมรรถนะของระบบบัตรเครดิตทั้งในด้านซอฟต์แวร์และฮาร์ดแวร์ อีกทั้งยังสะท้อนถึงความจำเป็นในการนำเทคโนโลยีสมัยใหม่เข้ามาช่วยวิเคราะห์ข้อมูลจากแฟ้มล็อก (log files) เพื่อหาสาเหตุและแก้ไขปัญหาคอมพิวเตอร์ที่ไม่เสถียรที่เกิดขึ้นอย่างมีประสิทธิภาพ

อย่างไรก็ตาม แม้จะมีการบริหารจัดการระบบบัตรเครดิตทั้งด้านซอฟต์แวร์และฮาร์ดแวร์ แต่ยังคงพบปัญหาคอมพิวเตอร์ที่ไม่เสถียรของระบบ โดยเฉพาะในช่วงเทศกาลที่มีการทำธุรกรรมจำนวนมาก ระบบมักหยุดชะงักและไม่สามารถให้บริการได้ตามปกติ จากข้อมูลสถิติพบว่าช่วงเวลาที่มีระบบเทคโนโลยีสารสนเทศขัดข้อง (downtime) ส่งผลกระทบต่อการทำงานของธนาคารพาณิชย์ มีค่าเฉลี่ยสูงถึงครึ่งละ 2 ชั่วโมง 15 นาที [5] ปัญหานี้ส่วนหนึ่งเกิดจากข้อจำกัดของเครื่องมือเฝ้าตรวจสอบระบบ (monitoring tool) ซึ่งยังไม่สามารถตอบสนองได้ทันทั่วถึง เนื่องจากเครื่องมือที่ใช้ในปัจจุบันมุ่งเน้นการตรวจสอบสถานะของอุปกรณ์โดยรวม เช่น การทำงานของซีพียู หน่วยความจำ และพื้นที่จัดเก็บข้อมูล แต่ไม่ได้วิเคราะห์เชิงลึกจากแฟ้มล็อก [6] เพื่อคาดการณ์ความผิดปกติได้ล่วงหน้า ส่งผลให้เกิดระบบเทคโนโลยีขัดข้องเป็นเวลายาวนานและกระทบต่อผู้ใช้บริการในวงกว้าง

ดังนั้นหนึ่งในแนวทางสำคัญในการแก้ไขปัญหาที่กล่าวมาข้างต้น คือการประยุกต์ใช้แนวคิดการเรียนรู้ของเครื่อง ซึ่งปัจจุบันเริ่มมีการประยุกต์ใช้การเรียนรู้ของเครื่องมาใช้ในการงานด้านอุตสาหกรรม การเงินและการธนาคารมากขึ้น [7] เพื่อเพิ่มศักยภาพในการตรวจสอบและคาดการณ์ความผิดปกติ

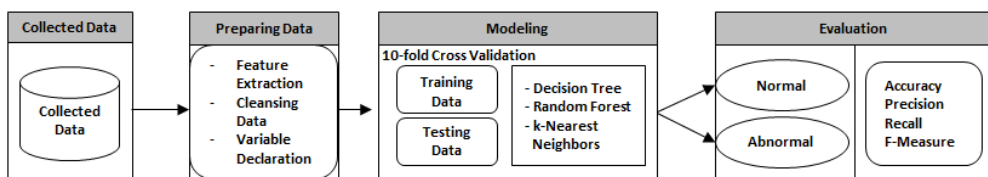
ในระบบทางการเงิน โดยเทคนิคการเรียนรู้ของเครื่องสามารถนำมาวิเคราะห์ข้อมูลจากบริการทางการเงินที่หลากหลาย [8] ช่วยให้การวิเคราะห์และการคาดการณ์มีความแม่นยำมากขึ้น ตลอดจนสนับสนุนการวางแผนและการตัดสินใจได้อย่างมีประสิทธิภาพ [9] จากประเด็นปัญหาด้านสมรรถนะของระบบบัตรเครดิตที่ยังขาดเสถียรภาพ และศักยภาพของการเรียนรู้ของเครื่องที่สามารถช่วยแก้ไขปัญหานี้ได้ งานวิจัยนี้จึงมุ่งเน้นการบูรณาการทั้งสองแนวคิดเข้าด้วยกัน โดยนำเสนอการพัฒนาแบบจำลองเพื่อตรวจสอบความผิดปกติของระบบบัตรเครดิตในเครือข่ายอีดีซี ด้วยวิธีการเรียนรู้ของเครื่อง เพื่อปรับปรุงสมรรถนะของระบบให้มีความเสถียรเพิ่มขึ้น [10] รวมถึงช่วยลดโอกาสการหยุดชะงักของระบบ ซึ่งจะนำไปสู่การสร้างเชื่อมั่นต่อผู้ใช้บริการได้อย่างมีประสิทธิภาพ ทั้งนี้ อัลกอริทึมการเรียนรู้ของเครื่องที่นำมาประยุกต์ใช้ประกอบด้วย ต้นไม้ตัดสินใจ (decision tree) ป่าสุ่ม (random forest) และเพื่อนบ้านใกล้ที่สุด (K-nearest neighbors) โดยจะทำการเปรียบเทียบและประเมินประสิทธิภาพของแบบจำลองทั้ง 3 อัลกอริทึม ด้วยค่าความถูกต้อง (accuracy) ค่าความแม่นยำ (precision) ค่าความครบถ้วน (recall) และค่าประสิทธิภาพโดยรวม (F-measure) [11, 12]

วัตถุประสงค์

1. เพื่อเปรียบเทียบและประเมินประสิทธิภาพแบบจำลองการตรวจสอบความผิดปกติของระบบบัตรเครดิตในเครือข่ายอีดีซี ด้วยวิธีการเรียนรู้ของเครื่อง
2. เพื่อหาตัวแบบจำลองที่เหมาะสมสำหรับนำไปพัฒนาเป็นเครื่องมือที่ใช้ในการตรวจสอบความผิดปกติของระบบบัตรเครดิตในเครือข่ายอีดีซี ด้วยวิธีการเรียนรู้ของเครื่อง

วิธีดำเนินการวิจัย

ขั้นตอนวิธีดำเนินการวิจัยประกอบด้วย 4 ขั้นตอน ดังภาพที่ 1



ภาพที่ 1 ขั้นตอนการดำเนินการวิจัย

1. การรวบรวมข้อมูล (collected data)

ข้อมูลที่ใช้ในการวิจัยครั้งนี้เป็นข้อมูลแฟ้มล็อกที่ได้จากอุปกรณ์เครือข่ายของระบบบัตรเครดิตในเครือข่ายอีดีซี รวมระยะเวลาทั้งสิ้น 5,014 ชั่วโมง ซึ่งประกอบไปด้วย 7 คุณสมบัติ ดังนี้ (1) จำนวนรายการเข้า (2) จำนวนรายการออก (3) ขนาดของข้อมูลขาเข้า (4) ขนาดของข้อมูลขาออก (5) จำนวนการเชื่อมต่อสำเร็จ (6) จำนวนการเชื่อมต่อไม่สำเร็จ (7) จำนวนการเชื่อมต่อทั้งหมด ข้อมูลดังกล่าวถูกผสานเข้ากับข้อมูลเหตุขัดข้องของระบบที่เกิดขึ้นในช่วงเวลาเดียวกัน จากนั้นได้ทำการติดป้ายกำกับสถานะการทำงานของระบบ (labeling) ให้กับชุดข้อมูล โดยแบ่งออกเป็น 2 กลุ่ม ได้แก่ สถานะปกติ

(normal) จำนวน 4,048 ชั่วโมง หรือคิดเป็น 80.70% และสถานะการทำงานผิดปกติ (abnormal) จำนวน 963 ชั่วโมง หรือคิดเป็น 19.30% หลังผ่านกระบวนการคัดเลือกและทำความสะอาดข้อมูล ได้ข้อมูลทั้งสิ้น 5,011 แถว ที่พร้อมใช้เป็นข้อมูลอ้างอิงในการพัฒนาแบบจำลองการตรวจสอบความผิดปกติ

2. การเตรียมข้อมูล (preparing data)

2.1 การคัดเลือกข้อมูล (feature extraction) จากข้อมูลที่ได้จากแฟ้มล็อกของอุปกรณ์เครือข่ายผานกับการติดป้ายกำกับสถานะการทำงาน พบว่าทุกคุณลักษณะมีผลกับสถานะการทำงานของระบบ (correlation analysis) ซึ่งประกอบด้วย 7 คุณลักษณะกับ 1 เป้าหมาย (target) โดยรายละเอียดคุณลักษณะที่นำมาสร้างตัวแบบจำลองที่ในงานวิจัย แสดงได้ดังตารางที่ 1

ตารางที่ 1 รายละเอียดคุณลักษณะที่นำมาสร้างแบบจำลอง

ลำดับ	ชื่อคุณลักษณะ	ชนิดตัวแปร	คำอธิบาย
1	จำนวนรายการเข้า	Integer	จำนวนรายการข้อมูลที่ได้รับจากระบบประมวลผล
2	จำนวนรายการออก	Integer	จำนวนรายการข้อมูลที่ส่งไปยังจากระบบประมวลผล
3	ขนาดของข้อมูลขาเข้า	Integer	ขนาดของข้อมูลที่ได้รับจากระบบประมวลผล
4	ขนาดของข้อมูลขาออก	Integer	ขนาดของข้อมูลที่ส่งไปยังระบบประมวลผล
5	จำนวนการเชื่อมต่อสำเร็จ	Integer	จำนวนการเชื่อมต่อกับระบบประมวลผลที่สำเร็จ
6	จำนวนการเชื่อมต่อไม่สำเร็จ	Integer	จำนวนการเชื่อมต่อกับระบบประมวลผลที่ไม่สำเร็จ
7	จำนวนการเชื่อมต่อทั้งหมด	Integer	จำนวนการเชื่อมต่อทั้งหมด
8	สถานะการทำงานของระบบ (เป้าหมาย)	Binomial	สถานะการทำงานของระบบ: การทำงานปกติ (normal) หรือ การทำงานผิดปกติ (abnormal)

2.2 การทำความสะอาดข้อมูล (cleansing data) จากการตรวจสอบข้อมูลแฟ้มล็อกพบว่า มีบางแถวที่ประกอบด้วยค่าที่ขาดหาย (missing values) ซึ่งอาจส่งผลกระทบต่อความถูกต้องและความสมบูรณ์ของแบบจำลองที่พัฒนา จากข้อมูลที่รวบรวมได้พบว่ามี 3 แถว ที่เป็นพบความไม่สมบูรณ์ จึงทำการตัดแถวนั้นออกเพื่อรักษาคุณภาพของชุดข้อมูลให้มีความสมบูรณ์ คงเหลือข้อมูล 5,011 แถว ที่ใช้สำหรับการนำไปประมวลผลและวิเคราะห์ต่อไป

2.3 การกำหนดค่าตัวแปร (variable declaration) ข้อมูลที่ได้จากแฟ้มล็อกเป็นข้อมูลเชิงตัวเลข (integer) ในขณะที่ข้อมูลที่ทำติดป้ายกำกับสถานะการทำงานของระบบ (label) ถูกกำหนดให้เป็นตัวแปรแบบทวิภาค (binominal variable) โดยแบ่งออกเป็น 2 ค่า ได้แก่ สถานะการทำงานปกติ และสถานะการทำงานผิดปกติ ทั้งนี้เพื่อให้สามารถนำไปใช้เป็นตัวแปรเป้าหมาย (target variable) ในกระบวนการจำแนกประเภทได้อย่างเหมาะสม

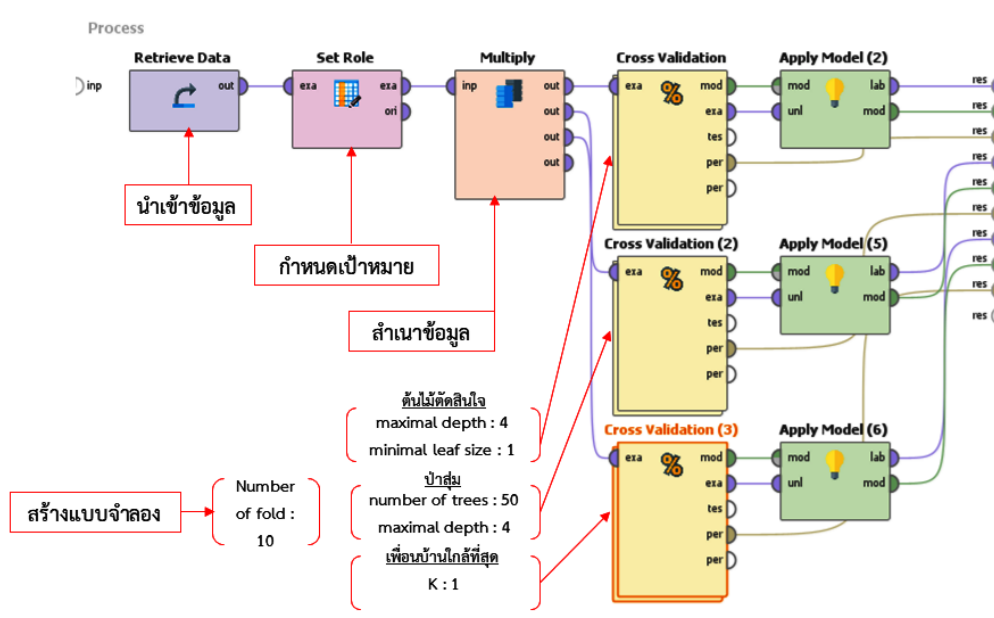
3. การพัฒนาแบบจำลอง (modeling) กระบวนการพัฒนาแบบจำลองประกอบด้วย 2 กระบวนการย่อย ดังนี้

3.1 การออกแบบการสร้างแบบจำลอง งานวิจัยนี้ได้ออกแบบการสร้างแบบจำลองโดยการแบ่งข้อมูลใช้สำหรับฝึกและสำหรับทดสอบ 9:1 ซึ่งใช้หลักการทดสอบแบบไขว้ทบ 10 ส่วน (10-fold cross validation) ทำให้ข้อมูลทุกส่วนจะถูกสลับเป็นข้อมูลสำหรับฝึกและข้อมูลสำหรับทดสอบ วนจนครบ 10 รอบ ซึ่ง 10-fold เป็นค่าที่เหมาะสมกับจำนวนข้อมูลในงานวิจัยนี้ ข้อดีของเทคนิคนี้คือ ข้อมูลทุกตัวจะถูกใช้ในการฝึกและทดสอบ และยังช่วยลดความเอนเอียงของข้อมูลได้ดี [13] สามารถอธิบายการทำงานของ การทดสอบแบบไขว้ทบ 10 ส่วน ได้ดังภาพที่ 2

	training set									testing set
รอบที่ 1	2	3	4	5	6	7	8	9	10	1
รอบที่ 2	1	3	4	5	6	7	8	9	10	2
รอบที่ 3	1	2	4	5	6	7	8	9	10	3
รอบที่ 4	1	2	3	5	6	7	8	9	10	4
รอบที่ 5	1	2	3	4	6	7	8	9	10	5
รอบที่ 6	1	2	3	4	5	7	8	9	10	6
รอบที่ 7	1	2	3	4	5	6	8	9	10	7
รอบที่ 8	1	2	3	4	5	6	7	9	10	8
รอบที่ 9	1	2	3	4	5	6	7	8	10	9
รอบที่ 10	1	2	3	4	5	6	7	8	9	10

ภาพที่ 2 การแบ่งข้อมูลแบบ 10-fold cross validation

3.2 การพัฒนาแบบจำลอง งานวิจัยนี้ได้ประยุกต์ใช้อัลกอริทึมการเรียนรู้ของเครื่องแบบมีผู้สอน ได้แก่ อัลกอริทึมต้นไม้ตัดสินใจ อัลกอริทึมป่าสุ่ม และอัลกอริทึมเพื่อนบ้านใกล้เคียงที่สุด ในการพัฒนาแบบจำลองสำหรับตรวจสอบความผิดปกติของธุรกรรมบัตรเครดิตในเครือข่ายอีดีซี โดยดำเนินการผ่านโปรแกรม RapidMiner ทั้ง 3 อัลกอริทึม แสดงได้ดังภาพที่ 3



ภาพที่ 3 การเชื่อมต่อเพื่อหาค่าประสิทธิภาพแบบจำลองทั้ง 3 อัลกอริทึม ด้วยโปรแกรม RapidMiner

ต้นไม้ตัดสินใจ เป็นอัลกอริทึมการเรียนรู้ของเครื่องแบบมีผู้สอน (supervise learning) ที่นิยมใช้ในงานจำแนกประเภท (classification) ซึ่งเข้าใจง่ายได้ผลดีและเป็นรากฐานของป่าสุ่ม [14] โดยหลักการทำงานจะทำการแบ่งข้อมูลออกเป็นกลุ่มย่อยตามเงื่อนไขที่กำหนด ผ่านการตัดสินใจในลักษณะ “ถ้า-แล้ว” (If-Then) ซ้ำไปเรื่อย ๆ จนได้ผลลัพธ์สุดท้าย การสร้างต้นไม้ตัดสินใจอาศัย ฟังก์ชันวัตถุประสงค์ (objective function) เช่น ค่าเอนโทรปี (entropy) หรือค่าดัชนีจีนิ (Gini index) ซึ่งสามารถอธิบายได้ดังสมการที่ 1

$$Gini(t) = 1 - \sum_{i=1}^n p_i^2 \quad (1)$$

โดยที่

t คือ โหนดที่ต้องการวัดค่า Gini

n คือ จำนวนคลาสทั้งหมดในโหนดนั้น

Pi คือ สัดส่วนของข้อมูลในโหนดที่อยู่ในคลาสที่ i

(จำนวนสมาชิกคลาส i หาค่าด้วยจำนวนสมาชิกทั้งหมดในโหนด)

การพัฒนาแบบจำลองโดยใช้อัลกอริทึมต้นไม้ตัดสินใจมีการกำหนดค่าพารามิเตอร์จำนวน ลำดับชั้น (maximal depth) และจำนวนโหนด (minimal leaf size) จากข้อมูลที่น่าสนใจ พบว่า ค่าพารามิเตอร์ที่เหมาะสมคือจำนวนลำดับชั้นเท่ากับ 4 และจำนวนโหนดเท่ากับ 1 โดยได้ผลการพัฒนาแบบจำลอง แสดงดังตารางที่ 2

ตารางที่ 2 ผลการพัฒนาแบบจำลองโดยใช้อัลกอริทึมต้นไม้ตัดสินใจ

จำนวน ลำดับชั้น	จำนวน โหนด	ค่าความถูกต้อง (ระบบไม่ปกติ)	ค่าความแม่นยำ (ระบบไม่ปกติ)	ค่าความครบถ้วน (ระบบไม่ปกติ)	ค่าประสิทธิภาพ โดยรวม
1	1	80.78	0	0	0
2	1	99.88	99.90	99.48	99.69
3	1	99.84	99.69	99.48	99.58
4	1	99.94	99.79	99.90	99.84
5	1	99.94	99.79	99.90	99.84
6	1	99.94	99.79	99.90	99.84

ป่าสุ่ม เป็นอัลกอริทึมที่พัฒนาต่อยอดจากอัลกอริทึมต้นไม้ตัดสินใจ โดยเป็นการรวมหลายต้นไม้ตัดสินใจ (ensemble learning) พบว่ามีหลาย ๆ งานวิจัยใช้อัลกอริทึมป่าสุ่มในการจำแนกข้อมูลแล้วได้ผลประสิทธิภาพสูง [15] ป่าสุ่มใช้วิธีการทดสอบกับข้อมูลชุดเดียวกันหลาย ๆ ครั้ง (instance) ผ่านการสุ่มตัวอย่างแบบมีการแทนที่ (bootstrap sampling) เพื่อสร้างชุดข้อมูลสำหรับการฝึกในแต่ละต้นไม้ ทั้งนี้ในการสร้างต้นไม้แต่ละต้นจะสุ่มเลือกเฉพาะคุณลักษณะบางส่วน (random subset of features) มาใช้ในการแบ่งข้อมูลที่แต่ละโหนด ส่งผลให้ต้นไม้แต่ละต้นมีความแตกต่างกัน สำหรับการทำนายผล ระบบจะรวมผลลัพธ์จากต้นไม้ทั้งหมดเข้าด้วยกัน และในการจำแนกประเภทจะใช้วิธีการโหวตเสียงข้างมาก (majority voting) เพื่อเลือกคลาสที่มีการโหวตสูงสุด

การพัฒนาแบบจำลองโดยใช้อัลกอริทึมป่าสุ่ม ผู้วิจัยได้กำหนดค่าพารามิเตอร์จำนวนลำดับชั้น (maximal depth) เท่ากับ 4 ซึ่งเป็นค่าที่ให้ผลลัพธ์ที่ดีที่สุดของอัลกอริทึมเป็นค่าเริ่มต้น พบว่าจำนวนต้นไม้ที่ให้ผลดีที่สุดเท่ากับ 50 ผลการพัฒนาแบบจำลอง แสดงดังตารางที่ 3

ตารางที่ 3 ผลการพัฒนาแบบจำลองโดยใช้อัลกอริทึมป่าสุ่ม

จำนวน ต้นไม้	จำนวน ลำดับชั้น (ระบบไม่ปกติ)	ค่าความ ถูกต้อง	ค่าความ แม่นยำ (ระบบไม่ปกติ)	ค่าความ ครบถ้วน (ระบบไม่ปกติ)	ค่า ประสิทธิภาพ โดยรวม
25	4	99.92	99.79	99.79	99.79
50	4	99.96	99.90	99.90	99.90
75	4	99.94	99.79	99.90	99.84
100	4	99.94	99.79	99.90	99.84
125	4	99.94	99.79	99.90	99.84
150	4	99.94	99.79	99.90	99.84

เพื่อนบ้านใกล้ที่สุด เป็นอัลกอริทึมการเรียนรู้ของเครื่องแบบมีผู้สอนในกลุ่มการจำแนกประเภท ซึ่งให้ค่าประสิทธิภาพสูงเมื่อใช้กับเทคนิคการทดสอบแบบไขว้ทบ (k-fold cross validation) [16] โดยหลักการทำงานคือการหาตัวอย่างข้อมูลที่มีความใกล้เคียงกับข้อมูลใหม่มากที่สุดจำนวน k ตัวอย่าง (neighbors) แล้วนำข้อมูลเหล่านั้นมาใช้ในการตัดสินใจเพื่อให้ได้ผลลัพธ์ กระบวนการคำนวณอาศัยการวัดระยะทางระหว่างข้อมูลใหม่กับข้อมูลทั้งหมดในชุดข้อมูลฝึก (training data) จากนั้นเลือกตัวอย่างที่ใกล้เคียงที่สุดจำนวน k ตัวอย่าง และในการจำแนกประเภทจะนับจำนวนคลาสของ k ตัวอย่างนั้น ก่อนที่จะเลือกคลาสที่ปรากฏมากที่สุดเป็นผลลัพธ์สุดท้าย สามารถอธิบายได้ดังสมการที่ 2

$$dist(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (2)$$

โดยที่

$dist(x, y)$ คือ ระยะห่างระหว่างข้อมูลตัวอย่าง x กับตัวอย่าง

n คือ จำนวนคุณลักษณะทั้งหมดของข้อมูลตัวอย่าง

x คือ คุณลักษณะตัวที่ i ของข้อมูลตัวอย่าง x และ

y คือ คุณลักษณะตัวที่ i ของข้อมูลตัวอย่าง y

การพัฒนาแบบจำลองโดยใช้อัลกอริทึมป่าสุ่มมีการกำหนดค่าพารามิเตอร์จำนวน k อ้างอิงจากงานวิจัย [16] โดยได้ผลการพัฒนาแบบจำลอง ดังตารางที่ 4

ตารางที่ 4 ผลการพัฒนาแบบจำลองโดยใช้อัลกอริทึมเพื่อนบ้านใกล้ที่สุด

จำนวน k	Cross validation	ค่าความถูกต้อง (ระบบไม่ปกติ)	ค่าความ แม่นยำ (ระบบไม่ปกติ)	ค่าความ ครบถ้วน (ระบบไม่ปกติ)	ค่า ประสิทธิภาพ โดยรวม
1	10	99.92	99.79	99.79	99.79
3	10	99.86	99.69	99.58	99.64
5	10	99.90	99.79	99.69	99.74
7	10	99.84	99.58	99.58	99.58

4. การวัดประสิทธิภาพแบบจำลอง (evaluation)

การประเมินประสิทธิภาพของแบบจำลองในแต่ละอัลกอริทึมที่ได้ โดยการเปรียบเทียบค่าที่แบบจำลองทำนายกับผลลัพธ์ที่เกิดขึ้นจริง และแสดงในรูปแบบของตารางที่เรียกว่าเมทริกซ์สับสน (confusion matrix) ตารางดังกล่าวช่วยให้เห็นภาพรวมของความถูกต้องในการจำแนกแต่ละคลาส ตลอดจนสามารถระบุประเภทของข้อผิดพลาดที่เกิดขึ้นได้ โดยในงานวิจัยนี้กำหนดให้คลาสการทำงานผิดปกติ (abnormal) เป็นกลุ่มที่ผู้วิจัยให้ความสนใจ (positive case) สำหรับรูปแบบตารางเมทริกซ์สับสนแสดงได้ดังตารางที่ 5

ตารางที่ 5 เมทริกซ์สับสน

pred. \ true	true abnormal	true normal
pred. abnormal (Postive Case)	TP	FP
pred. normal	FN	TN

โดยที่

TP คือ ค่าที่โปรแกรมทำนายถูกต้อง (ข้อมูลการทำงานผิดปกติ ทำนายว่าการทำงานผิดปกติ)

TN คือ สิ่งที่โปรแกรมทำนายถูกต้อง (ข้อมูลการทำงานปกติ ทำนายว่าการทำงานปกติ)

FP คือ สิ่งที่โปรแกรมทำนายไม่ถูกต้อง (ข้อมูลการทำงานผิดปกติ ทำนายว่าการทำงานปกติ)

FN คือ สิ่งที่โปรแกรมทำนายไม่ถูกต้อง (ข้อมูลการทำงานปกติ ทำนายว่าการทำงานผิดปกติ)

ค่าความถูกต้อง (accuracy) เป็นการประเมินโดยดูภาพรวมว่ามีค่าความถูกต้องของคลาสการทำงานผิดปกติ (abnormal) อธิบายได้ดังสมการที่ 3

$$Accuracy (abnormal) = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

ค่าความแม่นยำ (precision) แบบจำลองที่มีค่าความแม่นยำสูง หมายถึง ผลการทำนายตรงกับค่าความเป็นจริงของคลาสการทำงานผิดปกติ (abnormal) อธิบายได้ดังสมการที่ 4

$$Precision (abnormal) = \frac{TP}{TP+FP} \quad (4)$$

ค่าความครบถ้วน (recall) มีลักษณะคล้ายกันกับค่าความแม่นยำ แต่จะหารด้วยค่าที่ทำนายที่เกิดขึ้นจริง โดยค่าความครบถ้วนของคลาสการทำงานผิดปกติ (abnormal) อธิบายได้ดังสมการที่ 5

$$Recall (abnormal) = \frac{TP}{TP+FN} \quad (5)$$

ค่าประสิทธิภาพโดยรวม (F-measure) เป็นค่าที่ใช้วัดความแม่นยำของโมเดล โดยเป็นการหาค่าเฉลี่ยแบบ harmonic mean ระหว่างค่าความแม่นยำ และค่าความครบถ้วน อธิบายได้ดังสมการที่ 6

$$F - measure = \frac{2 \times (Precision \times Recall)}{(Precision+Recall)} \quad (6)$$

ผลการวิจัย

ผลการวิจัยในการประเมินประสิทธิภาพการจำแนกข้อมูลการทำงานของระบบบัตรเครดิตในเครือข่ายอีดีซี ใช้ชุดข้อมูลจำนวน 5,011 ชุด ประกอบด้วยคุณลักษณะ (feature) จำนวน 8 ตัวแปร และคลาส (class) จำนวน 2 ประเภท ได้แก่ กลุ่มข้อมูลปกติ และกลุ่มข้อมูลไม่ปกติ โดยใช้วิธีการทดสอบแบบไขว้ทบ 10 ส่วน สำหรับผลการจำแนกข้อมูลทั้ง 3 อัลกอริทึม ด้วยค่าพารามิเตอร์ที่เหมาะสม ให้ผลลัพธ์ตามตารางที่ 6

ตารางที่ 6 ผลการจำแนกข้อมูลโดยใช้วิธีการทดสอบแบบไขว้ทบ 10 ส่วน (10-fold Cross Validation)

ต้นไม้ตัดสินใจ accuracy: 99.94%

	True abnormal	True normal	Class precision
Pred. abnormal	962	2	99.79%
Pred. normal	1	4,046	99.98%
Class recall	99.90%	99.95%	

ป่าสุ่ม accuracy: 99.96%

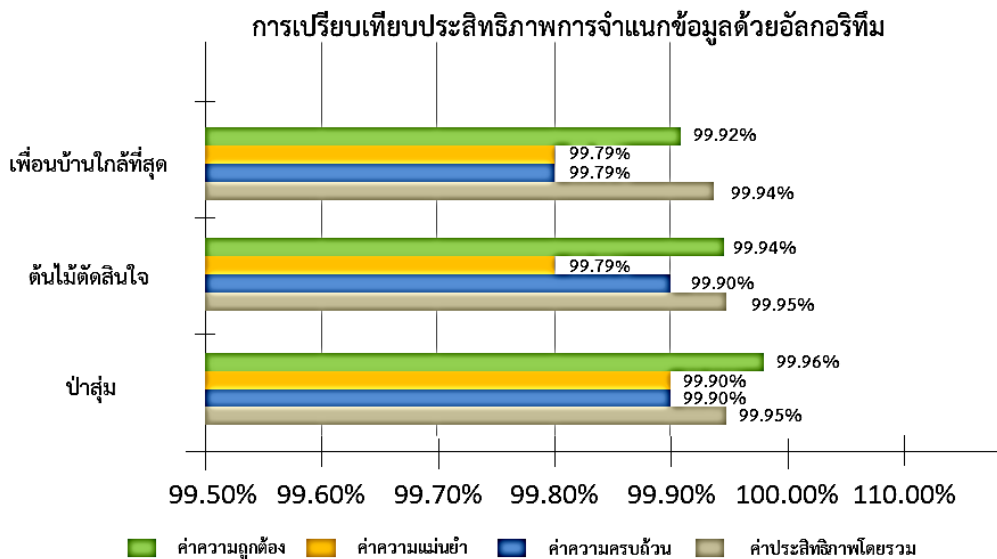
	True abnormal	True normal	Class precision
Pred. abnormal	962	1	99.90%
Pred. normal	1	4,047	99.98%
Class recall	99.90%	99.98%	

เพื่อนบ้านใกล้ที่สุด accuracy: 99.92%

	True abnormal	True normal	Class precision
Pred. abnormal	961	2	99.79%
Pred. normal	2	4,046	99.95%
Class recall	99.79%	99.95%	

จากตารางที่ 6 พบว่าแบบจำลองที่สร้างจากอัลกอริทึมป่าสุ่ม มีความถูกต้องในการทำนายว่าระบบผิดปกติมากที่สุดคือร้อยละ 99.96 ค่าความแม่นยำร้อยละ 99.90 ค่าความครบถ้วนร้อยละ 99.90 และค่าประสิทธิภาพโดยรวมร้อยละ 99.95 รองลงมาเป็นอัลกอริทึมต้นไม้ตัดสินใจ มีความถูกต้องในการทำนายว่าระบบผิดปกติร้อยละ 99.94 ค่าความแม่นยำร้อยละ 99.79 ค่าความครบถ้วนร้อยละ 99.90 และค่าประสิทธิภาพโดยรวมร้อยละ 99.95 และอัลกอริทึมเพื่อนบ้านใกล้ที่สุด มีความถูกต้องในการทำนายว่าระบบผิดปกติร้อยละ 99.92 ค่าความแม่นยำร้อยละ 99.79 ค่าความครบถ้วนร้อยละ 99.79 และค่าประสิทธิภาพโดยรวมร้อยละ 99.94 ตามลำดับ

เมื่อเปรียบเทียบทั้ง 3 อัลกอริทึม พบว่าอัลกอริทึมป่าสุ่มมีประสิทธิภาพโดยรวมสูงที่สุด โดยเฉพาะในด้านค่าความถูกต้องและค่าความแม่นยำที่คงที่และมีค่าสูงกว่าอัลกอริทึมอื่นเล็กน้อย ซึ่งสะท้อนถึงความสามารถในการจำแนกข้อมูลที่มีความเสถียรและน่าเชื่อถือมากกว่า ดังนั้นจึงสามารถสรุปผลตามแนวคิด predictive modeling ได้ดังภาพที่ 4



ภาพที่ 4 การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลด้วย 3 อัลกอริทึม

จากภาพที่ 4 ตัวแบบจำลองที่สร้างจากอัลกอริทึมป่าสุ่มที่มีความถูกต้อง ค่าความแม่นยำ ค่าความครบถ้วน และค่าประสิทธิภาพโดยรวมด้วยโปรแกรม RapidMiner โดยใช้วิธีการแบ่งข้อมูล 9:1 ด้วยการทดสอบแบบไขว้ทบ 10 รอบ

อภิปรายผลการวิจัย

การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลสถานะการทำงานของระบบทั้ง 3 อัลกอริทึม พบว่าอัลกอริทึมป่าสุ่ม ร่วมกับวิธีการทดสอบแบบไขว้ทบ 10 ส่วน ให้ผลลัพธ์ที่มีประสิทธิภาพสูงที่สุด โดยถ้าพิจารณาโครงสร้างของป่าสุ่มพบว่าโครงสร้างแบบต้นไม้หลายต้น (ensemble) และใช้ค่าเฉลี่ยจากต้นไม้หลายต้น ส่งผลให้มีความแม่นยำสูงและลดโอกาสการเกิด overfitting ได้ดี สอดคล้องกับงานวิจัย [15, 17] ที่พบว่าเทคนิคการจำแนกข้อมูลด้วยอัลกอริทึมป่าสุ่มให้ประสิทธิภาพในการจำแนกข้อมูลได้ดีกว่าทุกเทคนิคที่นำมาเปรียบเทียบ

สรุปผลการวิจัย

เอกสารอ้างอิง

- การตรวจสอบความผิดปกติของระบบบัตรเครดิตในเครือข่ายอิตีซีโดยใช้วิธีแบบปาล์ม

- [5] ธนาคารแห่งประเทศไทย. (2568). รายงานข้อมูลสถิติระบบเทคโนโลยีสารสนเทศขัดข้องกระทบต่อการให้บริการสำคัญของธนาคารพาณิชย์. [ออนไลน์], สืบค้นจาก <https://www.bot.or.th/th/statistics/system-failure-disclosure.html> (1 กรกฎาคม 2568).
- [6] ธนาธาร ภัทรกวิน, และประภาส จงสฤษดิ์วัฒนา. (2559). การตรวจสอบความผิดปกติของเครื่องจักรจากแฟ้มล็อกในกระบวนการผลิตฮาร์ดดิสก์ด้วยต้นไม้ตัดสินใจ. *วารสารไทยการวิจัยดำเนินงาน*, 4(2), 10-17.
- [7] Achary, R., & Shelke, J. C. (2023). Fraud detection in banking transaction using machine learning (pp. 221–226). In *International Conference on Intelligent and Innovative Technologies in Computing, Electronics (IITCEE)*. IEEE.
- [8] Hossain George, M. Z., Alam, M. K., & Hasan, M. T. (2023). Machine learning for fraud detection in digital banking: A systematic literature review. *ASRC Procedia: Global Perspectives in Science and Scholarship*, 3(1), 37–61.
- [9] สกฤดาภรณ์ ทองคำ, และนุวิทย์ วิวัฒนวัฒนา. (2565). การเรียนรู้ของเครื่องเพื่อการทำนายการผิดนัดชำระของลูกค้าหนี้บัตรเครดิต (หน้า 37–48). ใน *การประชุมวิชาการ วิทยาการข้อมูล ครั้งที่ 2*. มหาวิทยาลัยศรีนครินทรวิโรฒ.
- [10] Fathima, A., Shree Devi, G., & Gulzar, Z. (2023). Cyber security reinforcement through firewall log analysis and machine learning. *International Journal on Recent and Innovation Trends in Computing and Communication*, 11(10), 8574–8580.
- [11] ไพศาล สิมาลาเตา, และจรัญ แสนราช. (2562). การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลปัจจัยสนับสนุนการเรียนรู้การศึกษาระบบเปิดด้วยสื่อการสอนอิเล็กทรอนิกส์ในระดับอุดมศึกษา. *วารสารวิชาการครุศาสตร์อุตสาหกรรม พระจอมเกล้าพระนครเหนือ*. 10(1), 66–74.
- [12] จุฑาทิพย์ ทิพย์พูล, และนิเวศ จิระวิชิตชัย. (2559). การจำแนกจดหมายอิเล็กทรอนิกส์ที่เป็นสแปมโดยใช้เทคนิคเหมืองข้อมูล. *วารสารวิทยาศาสตร์และเทคโนโลยี มทร. รัตนบุรี*, 6(1), 102–109.
- [13] Darapureddy, N., Karatapu, N., & Battula, T. K. (2019). Research of machine learning algorithms using K-fold cross validation. *International Journal of Engineering and Advanced Technology*, 8(6S), 215–218.
- [14] ชิตพงษ์ กิตตินราดร. (2563). ต้นไม้ตัดสินใจ [ออนไลน์], สืบค้นจาก <https://guopai.github.io/ml-blog09.html> (1 กรกฎาคม 2568).
- [15] Al-Tarawneh, B. A., & Bani-Salameh, H. (2022). Classification of firewall logs actions using machine learning techniques and deep neural network [EasyChair Preprint No. 8317]. EasyChair.
- [16] Tarapitakwong, J., Chartrungruang, B., Somhom, S., & Tantranont, N. (2017). A classification model for predicting standard levels of OTOP's wood handicraft products by using the K-nearest neighbor. *International Journal of the Computer, the Internet and Management*, 25(2), 135–141.

- [17] พิรศุขม์ ทองพ่วง, และจรัญ แสนราช. (2567). การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลเพื่อทำนายการได้รับทุนการศึกษานักศึกษาระดับปริญญาตรี โดยใช้เทคนิควิธีการทำเหมืองข้อมูล. วารสารการพัฒนางานประจำสู่งานวิจัย, 8(2), 44–52.
- [18] Rahman, M. H., Islam, T., Rana, M. M., Tasnim, R., Mona, T. R., & Sakib, M. M. (2023). Machine learning approach on multiclass classification of internet firewall log files [Preprint]. arXiv.