วิธีการสร้างแบบจำลองเชิงทำนายพฤติกรรมการผิดเงื่อนไขการปล่อยชั่วคราวของศาลจากชุดข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการเรียนรู้ของเครื่อง

Predictive modeling approach of breach behaviors for provisional release to the court from an imbalanced dataset using machine learning techniques

วิทยา ปัญญา[1], วฤษาย์ ร่มสายหยุด[2*]
Wittaya Panya[1], Walisa Romsaiyud[2*]

Received: 11 August 2022 ; **Revised**: 13 September 2022 ; **Accepted**: 10 October 2022

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์ (1) เพื่อสร้างแบบจำลองสำหรับการทำนายพฤติกรรมการผิดเงื่อนไขการปล่อยชั่วคราวของศาลด้วยวิธีการสุ่มตัวอย่างสังเคราะห์ที่ปรับเปลี่ยนได้ (Adaptive Synthetic Sampling Approach: ADASYN) และ (2) เพื่อประเมินประสิทธิภาพของแบบจำลองการเรียนรู้ที่มีปัญหาชุดข้อมูลไม่สมดุล วิธีการวิจัยเป็นการแก้ปัญหาด้วยกระบวนการเรียนรู้ของเครื่อง (Machine Learning) ซึ่งกระบวนการนี้ประกอบด้วย 6 ขั้นตอน 1) เก็บรวบรวมข้อมูลจากศาลจังหวัดพะเยา ระหว่างเดือนมกราคม 2560 - พฤษภาคม 2565 จำนวนระเบียนทั้งหมด 2,577 ระเบียน และ 19 คุณลักษณะ จากการปล่อยชั่วคราวของศาลในคดีอาญา 2) การเตรียมข้อมูลโดยเปรียบเทียบวิธีการแก้ไขปัญหาข้อมูลไม่สมดุลจำนวน 4 วิธี ได้แก่ Random Oversampling, SMOTE, BorderlineSMOTE และ ADASYN เพื่อเรียนรู้จากชุดข้อมูลที่ไม่สมดุล ซึ่งมีข้อมูลกลุ่มมาก (majority) จำนวน 2,475 ระเบียน และข้อมูลกลุ่มน้อย (minority) จำนวน 102 ระเบียน หรือมีอัตราส่วนข้อมูลกลุ่มน้อยต่อข้อมูลกลุ่มมากคิดเป็น 1: 24.26 ผลการเปรียบเทียบพบว่าวิธี ADASYN เป็นวิธีที่ให้ประสิทธิภาพสูงสุด และใช้การเลือกคุณลักษณะแบบฝังตัว 3) สร้างแบบจำลองการจำแนกประเภทด้วยอัลกอริทึม Gradient Boosting Machines ที่มีประสิทธิภาพสูงสำหรับการเรียนรู้และทดสอบแบบจำลองเมื่อเปรียบเทียบกับอัลกอริทึม AdaBoost และ XGBoost 4) ประเมินประสิทธิภาพแบบจำลองด้วย 4 เมตริกหลัก คือค่าความถูกต้อง ค่าความแม่นยำ ค่าความครบถ้วน ค่าประสิทธิภาพโดยรวม 5) การปรับพารามิเตอร์ของแบบจำลองเพื่อหาค่าที่เหมาะสมที่สุด และ 6) การนำแบบจำลองไปใช้งาน สำหรับผลการประเมินประสิทธิภาพ มีค่าความถูกต้องคิดเป็นร้อยละ 97.44 ค่าความแม่นยำ 96.37 ค่าความครบถ้วน 98.39 และค่าประสิทธิภาพโดยรวม 97.46

คำสำคัญ: การสร้างแบบจำลองเชิงทำนาย ชุดข้อมูลที่ไม่สมดุล การเรียนรู้ของเครื่อง วิธีการสุ่มตัวอย่างสังเคราะห์ที่ปรับเปลี่ยนได้ Gradient Boosting Machines

Abstract

The purposes of this research were. - (1) to build a model for predicting the breach behaviors for provisional release for a provincial law court using the Adaptive Synthetic Sampling Approach and (2) to evaluate the performance of a model based on the imbalanced dataset problem. The research methodology was designed to solve the problem with the Machine Learning process. The process consists of 6 steps; - 1) data was collected from Phayao Provincial Court during the January 2017 - May 2022 comprising 2,577 records and 19 features from provisional releases to the court in crime cases, 2) data preparation by comparing methods to solve the imbalanced dataset problem with 4 methods; Random Oversampling, SMOTE, BorderlineSMOTE and ADASYN for learning data from the imbalanced dataset. The majority class had 2,475 examples and the minority class had 102 examples or a minority to majority

[1] แขนงวิชาเทคโนโลยีสารสนเทศและการสื่อสาร สาขาวิชาวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยสุโขทัยธรรมาธิราช นนทบุรี 11120
[2] รองศาสตราจารย์ สาขาวิชาวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยสุโขทัยธรรมาธิราช นนทบุรี 11120
[1] Information and Communication Technology, School of Science and Technology, Sukhothai Thammathirat Open University, Nonthaburi, 11120
[2] Associate Professor, School of Science and Technology, Sukhothai Thammathirat Open University, Nonthaburi, 11120
* Corresponding author E-mail: walisa.rom@stou.ac.th

ratio of 1: 24.26. Comparing results showed that the method with ADASYN had high performance and using feature selection with embedded approach, 3) build a classification model with the Gradient Boosting Machines algorithm of high efficiency for training and testing a model by comparing with AdaBoost and XGBoost, 4) performance evaluation of the model with four main metrics that were accuracy, precision, recall and F-measure, 5) parameter tuning for finding the optimal value and 6) implementation of the model. The experimental results showed that the model had high performance in predicting breach behaviors for provisional release for court with the measurement results of accuracy value 97.44 %, precision 96.37%, recall 98.39% and F-measure 97.46%.

Keywords: Predictive Modeling, Imbalanced Datasets, Machine Learning, Adaptive Synthetic Sampling Approach, Gradient Boosting Machines

บทนำ

ปัจจุบันมีการนำหลักการเรียนรู้ของเครื่อง (Machine Learning: ML) ซึ่งเป็นการทำให้คอมพิวเตอร์สามารถเรียนรู้จำรูปแบบ (pattern recognition) ของข้อมูล และสามารถทำการจำแนกข้อมูล (data classification) จัดกลุ่มข้อมูล (data clustering) หรือวิเคราะห์การถดถอย (regression analysis) เพื่อแก้ปัญหาได้อย่างอัตโนมัติจากข้อมูลที่ป้อนให้ (สำนักงานพัฒนารัฐบาลดิจิทัล, 2563) โดยการนำข้อมูลมาสร้างแบบจำลอง (model) สำหรับการทำงานในรูปแบบต่าง ๆ เช่น การวิเคราะห์เชิงทำนายสินค้า การตรวจจับการฉ้อโกงเงินธนาคาร การจัดกลุ่มลูกค้าสินเชื่อ การทำนายการลาออกกลางคัน การแนะนำสินค้า หรือการทำนายโรคอุบัติใหม่ ซึ่งแบบจำลองที่ดี ต้องมาจากข้อมูลที่ไม่เอนเอียง (non-bias) มีความสมดุล (balance) ของข้อมูลที่นำมาสร้างแบบจำลอง เพื่อให้เกิดการทำงานที่มีประสิทธิภาพ แต่ด้วยข้อมูลหลายประเภทที่เป็นข้อมูลไม่สมดุล (imbalanced data) กล่าวคือ ข้อมูลที่มีจำนวนรายการไม่เท่ากันหรือเอนเอียงไปทางคลาส (class) ใดคลาสหนึ่งมากเกินไป เช่น ข้อมูลคนฉ้อโกงเงินธนาคารต่อคนที่ไม่ฉ้อโกงเงินธนาคาร หรือข้อมูลจุดคาบน้ำมันรั่วในทะเล (จำนวนไม่มาก) ต่อพื้นที่ทะเล (จำนวนมากกว่า) จากงานวิจัยการพยากรณ์ผู้ป่วยเบาหวาน (วิชญ์วิสิฐ เกสรสิทธิ์ และคณะ, 2561) แก้ปัญหาข้อมูลไม่สมดุลโดยมีอัตราส่วนของข้อมูลระหว่างกลุ่มไม่กลับมารักษาซ้ำ กลับมารักษาซ้ำภายใน 30 วัน และกลับมารักษาซ้ำมากกว่า 30 วันเป็นร้อยละ 53.92: 11.16: 34.92 ตามลำดับ ด้วยวิธีสุ่มตัวอย่างสังเคราะห์เพิ่มของข้อมูลกลุ่มน้อย (Synthetic Minority Over-sampling Technique: SMOTE) (Chawla *et al.*, 2002) ร่วมกับขั้นตอนวิธี หรืออัลกอริทึมต้นไม้การตัดสินใจ (Decision Tree) สามารถจำแนกผู้ป่วยโรคเบาหวานได้ดีที่สุด และจากงานวิจัยของ Guan *et al.* (2021) ดำเนินการตรวจจับมัลแวร์บนระบบปฏิบัติการแอนดรอยด์โดยใช้การเรียนรู้ของเครื่องกับข้อมูลไม่สมดุล โดยใช้วิธีการปรับสมดุลให้ข้อมูลด้วยวิธีใช้วิธี SMOTE ทำให้มีประสิทธิภาพที่ดีขึ้น และการทำนายอัตราตอบรับการเสนอขายกรมธรรม์ของธนาคาร (กิตติภพ แซ่เตีย และจิรภัทร์ หยกรัตนศักดิ์, 2564) พบว่าการใช้วิธี SMOTE มีประสิทธิภาพเหมาะสมกับการแก้ปัญหาข้อมูลไม่สมดุลของการเสนอขายกรมธรรม์ เป็นต้น แต่วิธี SMOTE นั้นมีข้อจำกัด (Jiang *et al.*, 2021) คือ 1) การสุ่มตัวอย่างที่ทับซ้อนกัน 2) การสุ่มตัวอย่างทำให้มีสิ่งรบกวน (Noise) 3) เป็นการยากที่จะกำหนดจำนวนค่าคงที่ (k) ของขั้นตอนวิธีเพื่อนบ้านที่ใกล้ที่สุด และทำให้เกิดข้อผิดพลาดจากการมองไม่เห็นค่า k สำหรับกำหนดตำแหน่งด้วยวิธีเพื่อนบ้านที่ใกล้ที่สุดในการเลือกสำหรับสร้างตัวอย่างสังเคราะห์ นอกจากการใช้วิธี SMOTE แล้วยังมีการแก้ไขปัญหาข้อมูลที่ไม่สมดุลด้วยมีวิธี ADASYN ดังเช่น การจำแนกข้อมูลที่พัฒนาสำหรับชุดข้อมูลที่ไม่สมดุลของภาวะข้อเข่าเสื่อมในผู้สูงอายุ (พุทธิพร ธนธรรมเมธี และเยาวเรศ ศิริสถิตกุล, 2561) โดยใช้วิธีการสุ่มตัวอย่างสังเคราะห์ที่ปรับเปลี่ยนได้ (Adaptive Synthetic Sampling Approach: ADASYN) (He *et al.*, 2008) และวิธีสุ่มตัวอย่างสังเคราะห์เพิ่มของข้อมูลกลุ่มน้อย (Synthetic Minority Over-sampling Technique: SMOTE) และใช้วิธีการตรวจสอบแบบไขว้ 10 กลุ่ม (10-fold cross validation) ในการแบ่งเป็นชุดชุดข้อมูลฝึกสอน (training data) และชุดข้อมูลทดสอบ (testing data) ผลการทดสอบประสิทธิภาพพบว่าวิธี ADASYN ให้ค่าความถูกต้อง 97.31 % และมีงานวิจัยของ He และคณะได้ดำเนินการเปรียบเทียบประสิทธิภาพระหว่าง SMOTE และ ADASYN จากชุดข้อมูล 5 ชุดที่มีระดับความไม่สมดุลแตกต่างกัน พบว่า ADASYN มีประสิทธิภาพเหนือกว่า SMOTE จากงานวิจัยดังกล่าวแสดงให้เห็นว่าวิธีการ ADASYN เป็นวิธีแก้ไขปัญหาข้อมูลไม่สมดุลได้ดีกว่าวิธี SMOTE เนื่องจากวิธี ADASYN เป็นการปรับปรุงข้อจำกัดของ SMOTE ช่วยลดความเอนเอียง (bias) ที่เกิดจากความไม่สมดุลของคลาส และปรับเปลี่ยนขอบเขตการตัดสินใจจำแนกประเภทให้ดีขึ้น

ด้วยศาลจังหวัดพะเยาได้อนุญาตให้ผู้ต้องหาหรือจำเลยได้รับการปล่อยชั่วคราว (provisional release) จากการยื่นคำร้องต่อศาลเพื่อขอให้พิจารณาปล่อยผู้ต้องหา

หรือจำเลยชั่วคราว ซึ่งการปล่อยชั่วคราวมี 3 ประเภท คือ 1) การปล่อยชั่วคราวโดยไม่มีประกัน 2) การปล่อยชั่วคราวโดยมีประกัน และ 3) การปล่อยชั่วคราวโดยมีประกันและหลักประกัน ภายในระยะเวลาที่กำหนด (กองการต่างประเทศ สำนักงานศาลยุติธรรม, 2555) จากข้อมูลสถิติในโปรแกรมศาลชั้นต้นของศาลจังหวัดพะเยา ระหว่างเดือน มกราคม พ.ศ. 2560 - พฤษภาคม พ.ศ. 2565 มีการยื่นคำร้องขอปล่อยชั่วคราวในคดีอาญาและได้อนุญาตปล่อยชั่วคราว จำนวน 2,577 คำร้อง ซึ่งพบว่ามีผู้ที่ผิดสัญญาประกันหรือไม่ปฏิบัติตามเงื่อนไขของศาล หรือจำเลยหลบหนีหรือจะหลบหนีจำนวน 102 คำร้อง กรณีเช่นนี้ผู้ขอประกันอาจขอให้เจ้าพนักงานฝ่ายปกครองหรือตำรวจที่ใกล้ที่สุดจับจำเลย หรือศาลจะสั่งปรับผู้ถูกบังคับตามสัญญาประกันและยึดหลักประกันขายทอดตลาดหากไม่ชำระค่าปรับ ซึ่งส่งผลให้เพิ่มภาระหรือขั้นตอนในการปฏิบัติงานของเจ้าหน้าที่ และทำให้การพิจารณาคดีเกิดความล่าช้าได้ หากสามารถพยากรณ์พฤติกรรมที่อาจผิดสัญญาประกันหรือไม่ปฏิบัติตามเงื่อนไขของศาลได้ จะช่วยลดจำนวนผู้ถูกปล่อยชั่วคราวที่ผิดสัญญาประกันหรือผิดเงื่อนไขของศาล และลดขั้นตอนการปฏิบัติงานทำให้การพิจารณาคดีเป็นไปด้วยความรวดเร็วมากขึ้นได้ เมื่อพิจารณาข้อมูลที่ใช้ในจากปัญหาดังกล่าว งานวิจัยนี้จึงนำหลักการเรียนรู้ของเครื่องมาประยุกต์ใช้กับการวิเคราะห์เชิงทำนายการปล่อยชั่วคราวคดีอาญาของศาลจังหวัดพะเยา ตั้งแต่เดือนมกราคม พ.ศ. 2560 - พฤษภาคม พ.ศ. 2565 จำนวน 2,577 ระเบียน (record) และ 19 คุณลักษณะ (feature) มีข้อมูลผู้ถูกปล่อยชั่วคราวที่ไม่ผิดเงื่อนไขของศาล จำนวน 2,475 ระเบียน คิดเป็นร้อยละ 96.04 และมีผู้ที่ผิดผิดเงื่อนไขของศาล จำนวน 102 ระเบียน คิดเป็นร้อยละ 3.96 จะเห็นได้ว่าชุดข้อมูลดังกล่าวของศาลจังหวัดพะเยาประสบปัญหาข้อมูลที่ไม่สมดุล (imbalanced data) เกิดขึ้น เมื่อนำมาสร้างแบบจำลอง จะทำให้ได้แบบจำลองที่ไม่มีประสิทธิภาพ

ดังนั้น ในงานวิจัยนี้จึงได้นำวิธีแก้ปัญหาข้อมูลที่ไม่สมดุลด้วยวิธี ADASYN เพื่อปรับข้อมูลให้สมดุลร่วมกับผู้วิจัยได้นำเสนอการเพิ่มประสิทธิภาพของแบบจำลองโดยใช้วิธีการเลือกคุณลักษณะ (feature selection) แบบฝังตัว (embedded approach) จากนั้นสร้างแบบจำลองโดยได้เปรียบเทียบประสิทธิภาพของอัลกอริทึมในกลุ่มของ Boosting 3 อัลกอริทึม ได้แก่ 1) Adaptive Boosting: AdaBoost 2) Gradient Boosting Machines: GBM และ 3) eXtreme Gradient Boosting: XGBoost ซึ่งงานวิจัยนี้ต้องการแบบจำลองที่มีประสิทธิภาพสูง เนื่องจากเกี่ยวข้องกับชีวิตของคน และความเชื่อมั่นในการตัดสินของศาล ดังนั้นอัลกอริทึมในกลุ่มของ Boosting จึงถูกนำมาใช้ในการสร้างแบบจำลอง ที่มีจุดเด่นในการนำข้อมูลจากการเตรียมไว้ และทำการคัดเลือกคุณลักษณะ มาเป็นข้อมูลฝึกสอน (training data) ทำให้แบบจำลองมีประสิทธิภาพมากขึ้น และทำการปรับแต่งค่าพารามิเตอร์ที่เหมาะสม (Parameter Optimization) แบบอัตโนมัติ ของอัลกอริทึม เพื่อประเมินและเปรียบเทียบประสิทธิภาพของแบบจำลอง

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

เนื้อหาในส่วนนี้ขออธิบายทฤษฎีและงานวิจัยที่เกี่ยวข้อง จำนวน 3 เรื่อง ได้แก่ 1) ปัญหาความไม่สมดุลของข้อมูล (imbalanced data) 2) วิธีแก้ปัญหาความไม่สมดุลของข้อมูล 3) อัลกอริทึมแบบ Boosting

1. ปัญหาความไม่สมดุลของข้อมูล (imbalanced data) (Krawczyk, 2016) คือ สัดส่วนของจำนวนข้อมูลกลุ่มมาก (majority data) และข้อมูลกลุ่มน้อย (minority data) มีจำนวนที่แตกต่างกันมาก ซึ่งจะส่งผลต่อการสร้างแบบจำลองสำหรับการวิเคราะห์เชิงทำนายผล ทำให้แบบจำลองมีประสิทธิภาพต่ำ และมีความเอนเอียงในข้อมูลกลุ่มมาก ทำให้เกิดความไม่น่าเชื่อถือ โดยอัตราส่วนความไม่สมดุล (Imbalanced ratio: IR) สามารถหาได้จากสมการที่ 1 ดังนี้ (Zhua *et al.*, 2020)

$$\text{Imbalanced ratio (IR)} = \frac{\text{majority}}{\text{minority}} \qquad (1)$$

ตัวอย่างงานวิจัยนี้ มีกลุ่มข้อมูลที่ไม่ผิดเงื่อนไขหรือข้อมูลกลุ่มมาก (majority) จำนวน 2,475 ระเบียน และมีกลุ่มข้อมูลที่ผิดเงื่อนไขหรือข้อมูลกลุ่มน้อย (minority) จำนวน 102 ระเบียน สามารถคำนวณได้จากสมการ 1 จะได้ค่า IR = 24.26 ซึ่งอัตราส่วนข้อมูลกลุ่มน้อยต่อข้อมูลกลุ่มมากเป็น 1: 24.26

2. วิธีแก้ปัญหาความไม่สมดุลของข้อมูลด้วยการแก้ไขในระดับข้อมูลจะเป็นการปรับในขั้นของการเตรียมข้อมูล (data preparation) โดยการปรับให้ข้อมูลทั้งสองกลุ่มมีจำนวนที่ใกล้เคียงกัน ได้แก่ วิธีการสุ่ม (sampling methods) (Minh, 2018) คือ การสุ่มตัวอย่างข้อมูลเพื่อให้สมดุลหรือทำให้ข้อมูลแต่ละกลุ่มมีปริมาณที่สมดุลโดย แบ่งเป็น 2 กลุ่ม ได้แก่ 1) วิธีการสุ่มเพิ่ม (Oversampling) (Kulkarni *et al.*, 2020) เป็นวิธีการสุ่มนี้มีเป้าหมาย คือ การเพิ่มจำนวนตัวอย่างของคลาสส่วนน้อยเพื่อให้มีจำนวนเท่ากันหรือใกล้เคียงกับตัวอย่างของคลาสส่วนใหญ่ 2) วิธีการสุ่มลด (Undersampling) (Kulkarni *et al.*, 2020) เป็นวิธีการสุ่มตัวอย่างโดยลบตัวอย่างของคลาสส่วนใหญ่ในข้อมูลเพื่อทำให้ข้อมูลมีความสมดุลมากขึ้น แต่ข้อจำกัด คือ อาจสูญเสียข้อมูลสำคัญที่อาจเกิดขึ้น การลบเหตุการณ์ส่วนใหญ่ออกมากพอที่จะทำให้คลาสส่วนใหญ่มีขนาดเท่ากันหรือใกล้เคียงกับคลาสส่วนน้อย

ส่งผลให้สูญเสียข้อมูลที่สำคัญ เมื่อข้อมูลถูกลบออกโดยไม่ได้พิจารณาว่าเป็นอย่างไรและมีประโยชน์ต่อการวิเคราะห์เพียงใด ข้อเสียอีกประการคือกลุ่มตัวอย่างของคลาสส่วนใหญ่ที่เลือกอาจมีอคติ และผลของการวิเคราะห์อาจไม่ถูกต้อง จากงานวิจัยที่ได้ทำการตรวจจับการบุกรุกเป็นหัวข้อสำคัญในด้านความปลอดภัยทางไซเบอร์ และภัยคุกคามเครือข่ายทั่วไปที่มีข้อมูลไม่สมดุลอย่างมาก (Chen *et al.*, 2021) โดยเปรียบเทียบการแก้ไขด้วยวิธี SMOTE, ADASYN และ Random Undersampling พบว่า ADASYN มีประสิทธิภาพที่ดีที่สุด โดยมีค่า Precision 98.505%, Recall 93.606%, F1 95.303% และ AUC = 0.997 ดังนั้นงานวิจัยนี้จึงเลือกใช้วิธีการแก้ไขปัญหาข้อมูลไม่สมดุลด้วยวิธี Oversampling ซึ่งเป็นการรักษาความครบถ้วนของข้อมูลโดยไม่สูญเสียข้อมูลที่อาจเป็นข้อมูลสำคัญหรือมีผลต่อประสิทธิภาพของแบบจำลอง ซึ่งวิธีในกลุ่มของ Oversampling ที่นิยมใช้ในปัจจุบัน เช่น

2.1 วิธีการสุ่มตัวอย่างแบบเพิ่ม (Random Oversampling) (Kulkarni *et al.*, 2020) ในการสุ่มตัวอย่างมากเกินไป ตัวอย่างจากคลาสของกลุ่มน้อยจะถูกสุ่มเลือกสำหรับการทำซ้ำซึ่งส่งผลให้มีการกระจายคลาสที่สมดุล ในวิธีนี้ตัวอย่างของคลาสกลุ่มน้อยจะถูกสุ่มเลือก แต่มีข้อจำกัด คือการเพิ่มข้อมูลลักษณะนี้อาจจะทำให้เกิดปัญหาข้อมูลถูกรบกวนได้ง่าย

2.2 วิธีการสุ่มตัวอย่างสังเคราะห์ของข้อมูลกลุ่มน้อย (Synthetic Minority Over-sampling Technique: SMOTE) (Chawla *et al.*, 2002) เป็นวิธีการเพิ่มตัวอย่างข้อมูลโดยการสุ่มสร้างตัวอย่างข้อมูลขึ้นมาใหม่ ด้วยการนำตัวอย่างข้อมูลจากกลุ่มตัวอย่างที่มีจำนวนน้อยมาพิจารณาทีละตัวจนครบทุกตัว หลักการคือกำหนดจำนวนด้วยอัลกอริทึมเพื่อนบ้านที่ใกล้เคียงที่สุด (K-Nearest Neighbor: KNN) จำนวน k ตัว แล้วทำการสุ่มสร้าง ข้อมูลขึ้นมาใหม่ในพื้นที่ใดๆ บนทางที่เชื่อมโยงระหว่างจุดข้อมูลที่กำลังพิจารณาและจุดของข้อมูลเพื่อนบ้านที่ใกล้ที่สุด (K-nearest neighbor) แสดงดัง Figure 1

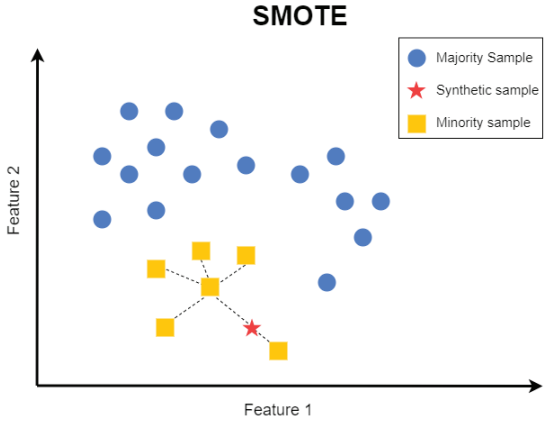

Figure 1 Synthetic Minority Over-sampling Technique (SMOTE)

จาก Figure 1 ประกอบด้วยรูปวงกลมแทนข้อมูลกลุ่มมาก รูปสี่เหลี่ยมแทนข้อมูลกลุ่มน้อย และรูปดาวเป็นข้อมูลที่สังเคราะห์ขึ้น

ซึ่งวิธี SMOTE มีข้อจำกัด (Jiang *et al.*, 2021) คือ 1) การสุ่มตัวอย่างที่ทับซ้อนกัน 2) การสุ่มตัวอย่างทำให้มีสิ่งรบกวน (noise) และ 3) เป็นการยากที่จะกำหนดจำนวนเพื่อนบ้านที่ใกล้ที่สุด และทำให้มองไม่เห็นเพื่อนบ้านที่ใกล้ที่สุดในการเลือกสำหรับสร้างตัวอย่างสังเคราะห์

2.3 วิธีการสุ่มตัวอย่างสังเคราะห์ของข้อมูลกลุ่มน้อยด้วยเส้นแนวเขต (Borderline-SMOTE) (Han *et al.*, 2005) เป็นวิธีการปรับปรุงจากวิธี SMOTE โดยการแก้ปัญหาในการเลือกตัวอย่างที่อยู่ใกล้แนวเขตของข้อมูลกลุ่มน้อยที่จำแนกผิด และสร้างเฉพาะตัวอย่างสังเคราะห์ที่ยากในการจำแนก อย่างไรก็ตามแม้จะเป็นวิธีที่ปรับปรุงจาก SMOTE แต่ยังมีข้อจำกัด คือ อาจสร้างตัวอย่างบางส่วนที่ทับซ้อนกับตัวอย่างเชิงลบและวิธีการสุ่มตัวอย่างมากเกินไปจะทำซ้ำทุกตัวอย่างที่เป็นบวกโดยไม่ต้องโดยคำนึงถึงปัจจัยอื่นๆ

2.4 วิธีการสุ่มตัวอย่างสังเคราะห์ที่ปรับเปลี่ยนได้ (Adaptive Synthetic Sampling Approach: ADASYN (He *et al.*, 2008) เป็นการสร้างตัวอย่างสังเคราะห์ (synthetic data) ที่ไม่จำเป็นต้องพิจารณาข้อมูลทุกตัวที่อยู่ในกลุ่มน้อย โดยใช้ค่าการแจกแจงแบบถ่วงน้ำหนัก (weight distribution) ของข้อมูลตัวอย่างในกลุ่มน้อย โดยการสังเคราะห์ข้อมูลซึ่งขึ้นอยู่กับความสำคัญของข้อมูลนั้นๆ ถ้าข้อมูลใดยากต่อการจำแนกก็จะให้ค่าของน้ำหนักข้อมูลนั้นมากและสังเคราะห์ข้อมูลชุดข้อมูลขึ้นมาในบริเวณนั้นๆ แสดงดัง Figure 2

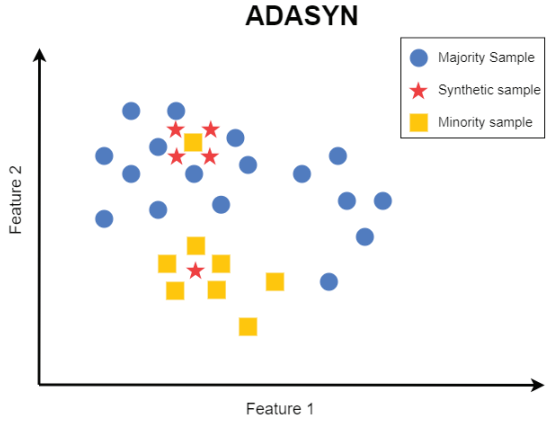

Figure 2 Adaptive Synthetic Sampling Approach (ADASYN)

จาก Figure 2 ประกอบด้วยรูปวงกลมแทนข้อมูลกลุ่มมาก รูปสี่เหลี่ยมแทนข้อมูลกลุ่มน้อย และรูปดาวเป็นข้อมูลที่สังเคราะห์ขึ้น ซึ่งเป็นวิธีที่ปรับปรุงการทำงานของวิธี SMOTE ให้ดีขึ้น ช่วยลดความเอนเอียงที่เกิดจากความไม่สมดุลของ

ข้อมูล และทำให้มีการปรับขอบเขตของการตัดสินใจในการจำแนกกลุ่มดีขึ้น โดยงานวิจัยนี้ใช้วิธีนี้ในการปรับสมดุลข้อมูล

3. อัลกอริทึมแบบ Boosting (Schapire & Freund, 2012) เป็นอัลกอริทึมการเรียนรู้ที่อ่อนแอ (Weak Learner) ซึ่งให้ตัวอย่างการฝึกอบรมที่มีป้ายกำกับ (labeled data) สร้างฐานหรือลักษณนามที่อ่อนแอ เป้าหมาย คือเพื่อปรับปรุงประสิทธิภาพของอัลกอริทึมการเรียนรู้ที่อ่อนแอโดยจะกำหนดน้ำหนักแต่ละรายการ แสดงการทำงานดัง Figure 3

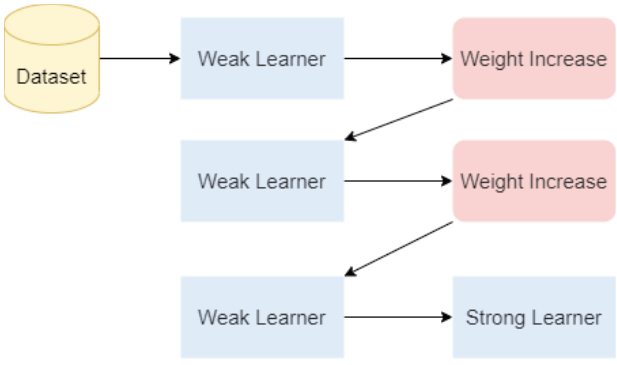

Figure 3 Boosting Algorithms

ซึ่งอัลกอริทึมที่จัดอยู่ในกลุ่มของ boosting ที่นิยมใช้ ได้แก่

3.1 อัลกอริทึม Adaptive Boosting: AdaBoost (Wu & Zhao, 2011) เป็นอัลกอริทึม Boosting แบบปรับตัวเองได้ซึ่งปรับปรุงประสิทธิภาพของตัวแยกประเภทที่อ่อนแอโดยการสร้างชุดของตัวแยกประเภทหลายตัว โดยปรับอัตโนมัติตามอัตราความผิดพลาดของอัลกอริทึมพื้นฐานในการฝึกผ่านการควบคุมแบบไดนามิกของน้ำหนักของแต่ละตัว

3.2 อัลกอริทึม Gradient Boosting Machines: GBM (Natekin & Knoll, 2013) เป็นวิธีที่พยายามจะสร้างแบบจำลองการทำนายผ่าน back-fitting และการถดถอย GBM เริ่มต้นด้วยการสร้างแบบจำลองเริ่มต้นและเรียนรู้ โดยเรียนรู้จากค่าความคลาดเคลื่อนสะสมที่เกิดจากการทำนายของตัวเรียนรู้ก่อนหน้า เพื่อให้ได้แบบจำลองที่มีความแม่นยำสูงสุด ซึ่งเป็นอัลกอริทึมที่มีประสิทธิภาพที่สุดในด้านการเรียนรู้ของเครื่อง เนื่องจาก GBM เป็นการเพิ่มประสิทธิภาพ จึงถูกใช้เพื่อลดข้อผิดพลาดความอคติหรือความเอนเอียง (Bias) ของแบบจำลองให้เหลือน้อยที่สุด

3.3 อัลกอริทึม eXtreme Gradient Boosting: XGBoost (Chen & Guestrin, 2016) เป็นเทคนิคที่พัฒนามาจาก Gradient Boosting ซึ่ง XGBoost เป็นแบบจำลองที่นำเอาต้นไม้ตัดสินใจมาฝึกสอนต่อกันหลายๆ ต้น โดยที่ต้นไม้ตัดสินใจแต่ละต้นจะเรียนรู้จากค่าความผิดพลาดของต้นก่อนหน้า ซึ่งทำให้ความแม่นยำในการทำนายจะมากขึ้นเรื่อยๆ เมื่อมีการเรียนรู้ของต้นไม้ตัดสินใจต่อเนื่องกันจนมีความลึกมากพอแบบจำลองจะหยุดเรียนรู้เมื่อไม่เหลือค่าความผิดพลาดจากต้นไม้ตัดสินใจต้นก่อนหน้าให้เรียนรู้แล้ว

วิธีดำเนินการวิจัย

วิธีการวิจัยครั้งนี้แบ่งออกเป็น 6 ขั้นตอน ประกอบด้วยขั้นตอนที่ 1) การเก็บข้อมูล 2) การเตรียมข้อมูล 3) สร้างแบบจำลอง 4) ประเมินประสิทธิภาพแบบจำลอง 5) การปรับพารามิเตอร์ และ 6) การนำไปใช้งาน โดยผู้วิจัยเลือกใช้เครื่องมือ Google Colaboratory (Colab) ด้วยภาษา Python สามารถแสดงภาพรวมของวิธีดำเนินการวิจัย ดัง Figure 4

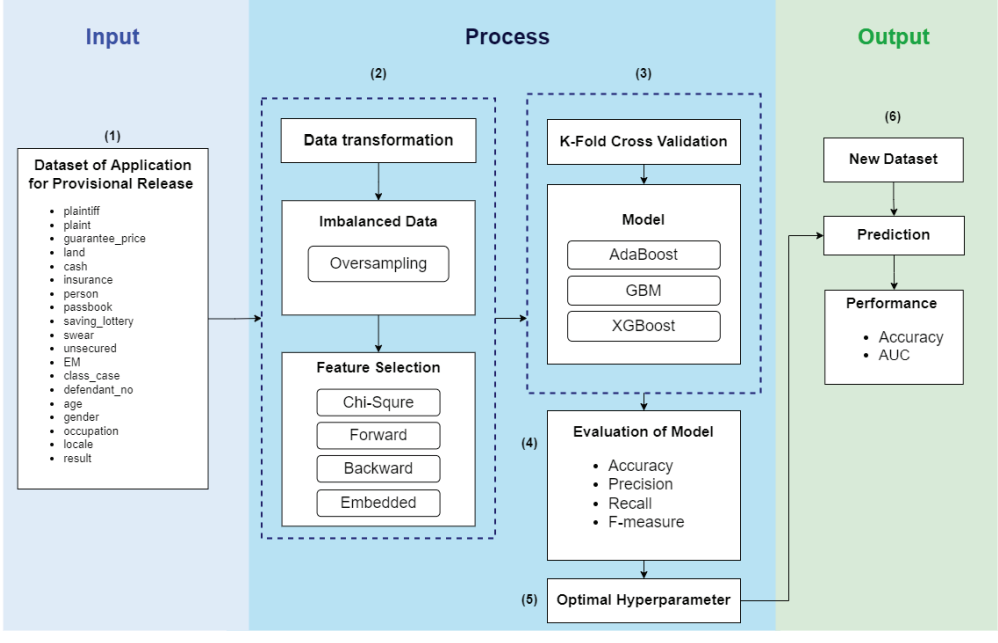

Figure 4 Overview of Research Methodology

1. การเก็บข้อมูล โดยใช้ข้อมูลผู้ถูกปล่อยชั่วคราวในคดีอาญาจากฐานข้อมูลโปรแกรมศาลชั้นต้นของศาลจังหวัดพะเยา ตั้งแต่เดือนมกราคม พ.ศ. 2560 - พฤษภาคม พ.ศ. 2565 จำนวน 2,577 ระเบียน และ 19 คุณลักษณะ ได้แก่ โจทก์ ข้อหา ราคาหลักประกัน หลักประกันที่ดิน หลักประกันเงินสด หลักประกันกรมธรรม์ หลักประกันบุคคล หลักประกันสมุดเงินฝาก หลักประกันสลากออมทรัพย์ สาบานตน ไม่มีหลักประกัน อุปกรณ์อิเล็กทรอนิกส์ติดตามตัว ชั้นของคดี ลำดับของจำเลย อายุ เพศ อาชีพ สถานที่เกิดเหตุ ผลการผิดเงื่อนไขหรือไม่ แสดงตัวอย่างชุดข้อมูล ดัง Figure 5

	plaintiff	plaint	guarantee_price	land	cash	insurance	person	passbook	saving_lottery	swear	unsecured	EM	class_case	defendant_no	age	gender	occupation	locale	result
0	1	68	0	0	0	0	0	0	0	0	1	0	พิจารณา	1	26	1	รับจ้าง	1	1
1	1	93	280050	1	1	0	0	0	0	0	0	0	พิจารณา	1	20	1	รับจ้าง	1	1
2	1	68	280050	1	1	0	0	0	0	0	0	0	พิจารณา	1	20	1	รับจ้าง	1	1
3	1	68	10000	0	1	0	0	0	0	0	0	0	พิจารณา	1	28	1	ไม่ทราบ	7	1
4	1	68	500480	1	1	0	0	0	0	0	0	0	พิจารณา	1	29	1	ไม่ทราบ	5	1
5	1	68	530460	1	0	0	0	0	0	0	0	0	พิจารณา	1	29	1	ไม่ทราบ	5	1
6	1	68	0	0	0	0	0	0	0	1	0	0	พิจารณา	1	30	1	รับจ้าง	5	0
7	1	64	0	0	0	0	0	0	0	1	0	0	พิจารณา	1	30	1	รับจ้าง	5	0
8	1	68	10000	0	1	0	0	0	0	0	0	0	พิจารณา	1	46	2	รับจ้าง	1	0
9	1	68	3000	0	1	0	0	0	0	0	0	0	พิจารณา	1	36	1	เกษตรกร	5	0

Figure 5 Example of Dataset

2. การเตรียมข้อมูล ประกอบด้วย

2.1 การแปลงข้อมูล เป็นการทำให้ข้อมูลอยู่ในรูปแบบที่พร้อมนำเอาไปใช้ในการวิเคราะห์ข้อมูลตามอัลกอริทึมที่เลือกใช้ ได้แก่ ทำการทำความสะอาดข้อมูลเพื่อตัดข้อมูลที่ไม่สอดคล้องกันออก การจัดกลุ่มของคุณลักษณะ รวมถึงข้อมูลที่มีค่าขาดหายไป (missing values)

2.2 แก้ไขปัญหาข้อมูลไม่สมดุล จากคุณลักษณะของชุดข้อมูลตาม Figure 5 มีคุณลักษณะที่จำแนกผลลัพธ์การผิดเงื่อนไขหรือไม่ผิดเงื่อนไข (result) พบว่ามีปัญหาข้อมูลไม่สมดุล (imbalanced data) คือ มีสัดส่วนของกลุ่มข้อมูลที่แตกต่างกัน ได้แก่ ข้อมูลที่เก็บเป็นตัวเลข 0 หมายถึง กลุ่มข้อมูลที่ไม่ผิดเงื่อนไขหรือข้อมูลกลุ่มมาก (majority) มีจำนวน 2,475 ระเบียน คิดเป็นร้อยละ 96.04 และข้อมูลที่เก็บเป็นตัวเลข 1 หมายถึง กลุ่มข้อมูลที่ผิดเงื่อนไขหรือข้อมูลกลุ่มน้อย (minority) มีจำนวน 102 ระเบียน คิดเป็นร้อยละ 3.96 โดยอัตราส่วนความไม่สมดุล (Imbalanced ratio: IR) คำนวณได้จากสมการที่ 1 จะได้ค่า IR = 24.26 ซึ่งอัตราส่วนข้อมูลกลุ่มน้อยต่อข้อมูลกลุ่มมากเป็น 1: 24.26 แสดงดัง Figure 6

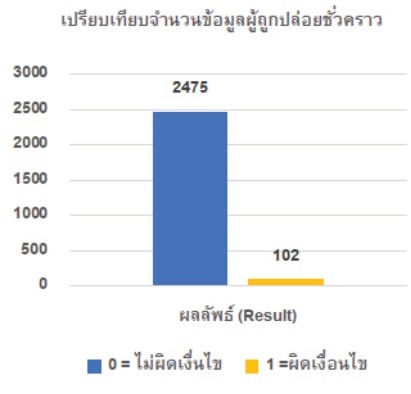

Figure 6 Imbalanced Data

ดังนั้น ผู้วิจัยจึงนำวิธีแก้ปัญหาความไม่สมดุลของข้อมูลโดยใช้อัลกอริทึม 3 อัลกอริทึม ได้แก่ 1) AdaBoost 2) GBM และ 3) XGBoost สำหรับสร้างแบบจำลองในเปรียบเทียบวิธีแก้ปัญหาความไม่สมดุลจำนวน 4 วิธี ได้แก่ 1) Random Oversampling 2) SMOTE 3) Borderline-SMOTE และ 4) ADASYN เพื่อประเมินประสิทธิภาพที่ดีที่สุดในการนำไปเรียนรู้ของแบบจำลอง โดยใช้เมทริกซ์ความสับสน (Confusion Matrix) (Kulkarni *et al.*, 2020) เป็นเครื่องมือประเมินที่ได้รับความนิยมอย่างมากในการแก้ปัญหาการจำแนกประเภท (classification) สามารถนำไปใช้กับการจำแนกไบนารีหรือปัญหาการจำแนกประเภทหลายคลาส ตัวอย่างของ Confusion Matrix สำหรับการจำแนกไบนารี แสดงดัง Figure 7

Confusion Matrix

	Predict: Positive (P)	Predict: Negative (N)
Actual: Positive	True Positive (TP)	False Negative (FN)
Actual: Negative	False Positive (FP)	True Negative (TN)

Figure 7 Confusion Matrix

โดย 1) TP เป็นจำนวนข้อมูลที่ทำนายถูกต้องในเชิงบวก (positive) หมายถึงข้อมูลที่ทำนายว่าไม่ผิดเงื่อนไขตรงกับความเป็นจริง 2) TN เป็นจำนวนข้อมูลที่ทำนายถูกต้องในเชิงลบ (negative) หมายถึงข้อมูลที่ทำนายว่าผิดเงื่อนไขตรงกับความเป็นจริง 3) FP เป็นจำนวนข้อมูลที่ทำนายผิดว่าอยู่ในเชิงบวก (positive) หมายถึงข้อมูลทำนายว่าไม่ผิดเงื่อนไขแต่ความเป็นจริงผิดเงื่อนไข 4) FN เป็นจำนวนข้อมูลที่ทำนายผิดว่าอยู่ในเชิงลบ (negative) หมายถึงข้อมูลทำนาย

ว่าผิดเงื่อนไขแต่ความเป็นจริงไม่ผิดเงื่อนไขซึ่งการประเมินประสิทธิภาพแบบจำลองคำนวณจากสมการ 2-7 ดังนี้

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

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

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

$$\text{F-measure} = \frac{2*(\text{Precision}*\text{Recall})}{(\text{Precision}+\text{Recall})} \quad (5)$$

$$\text{True Positive Rate (TPR)} = \frac{TP}{TP+FN} \quad (6)$$

$$\text{False Positive Rate (FPR)} = \frac{FP}{FP+TN} \quad (7)$$

ในงานวิจัยนี้จึงใช้การประเมินประสิทธิภาพแบบจำลอง โดยคำนวณค่า Accuracy Precision Recall และ F-measure ผลการประเมินประสิทธิภาพของแบบจำลองพบว่าวิธีการแก้ไขปัญหาข้อมูลไม่สมดุล ด้วยวิธี ADASYN มีประสิทธิภาพสูงสุด ซึ่งงานวิจัยนี้จึงเลือกวิธี ADASYN ไปใช้ในการแก้ไขปัญหาข้อมูลไม่สมดุล แสดงดัง Figure 8

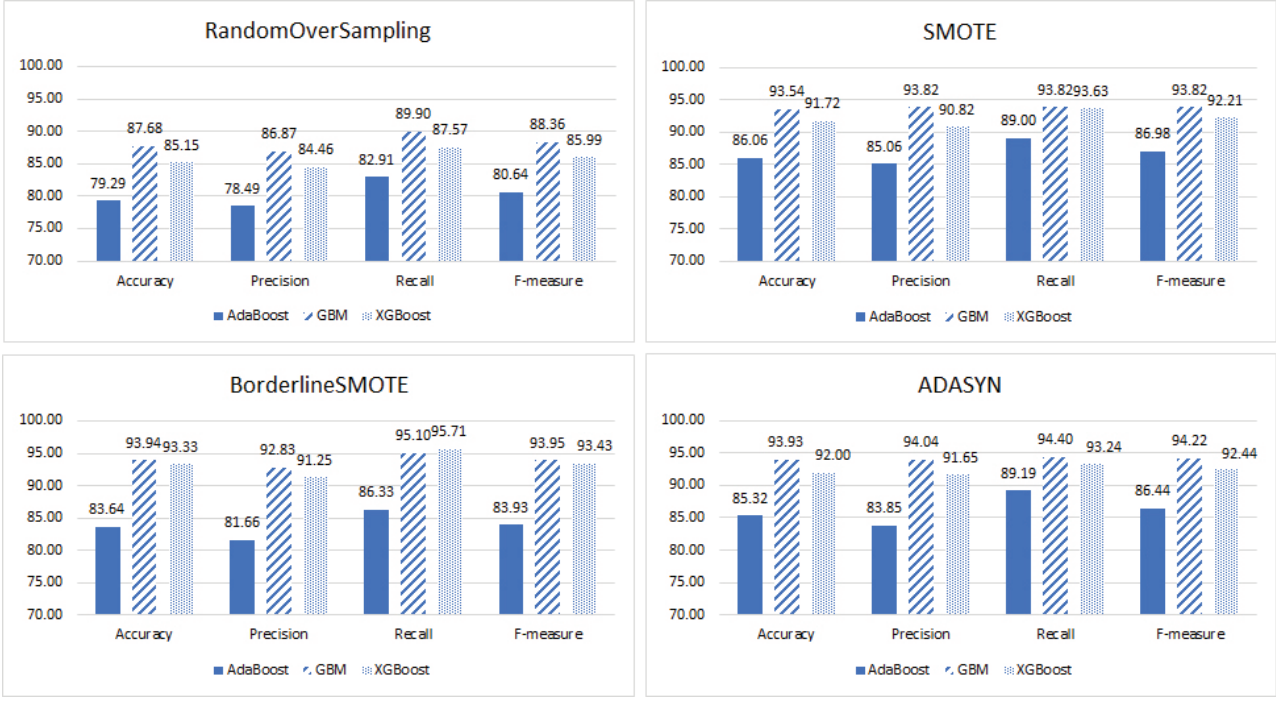

Figure 8 The results of evaluating the performance of a solution for modeling

2.3 การเลือกคุณลักษณะ (feature selection) (Gheyas & Smith, 2010) การเลือกคุณลักษณะเป็นวิธีการที่ช่วยลดจำนวนคุณลักษณะหรือแอททริบิวต์ (attribute) ซึ่งจะช่วยเพิ่มประสิทธิภาพและความแม่นยำของแบบจำลองในการจำแนกข้อมูล ซึ่งมี 3 ประเภท ได้แก่ 1) Filter เป็นการเลือกคุณสมบัติโดยไม่ขึ้นอยู่กับชนิดของคุณลักษณะที่ใช้ ข้อดี คือ เป็นเทคนิคที่คำนวณได้ง่าย รวดเร็ว และหลีกเลี่ยงการเกิด Overfitting เพราะไม่นำผลทดสอบมาพิจารณาร่วมด้วย ซึ่งคุณลักษณะที่ถูกคัดเลือกจะไม่มีอคติ (Bias) สำหรับข้อเสีย คือ คุณลักษณะที่ถูกคัดเลือกเป็นคุณลักษณะที่เป็นอิสระต่อกัน เพราะขั้นตอนการคำนวณค่าความสำคัญจะพิจารณาความสัมพันธ์เพียงด้านเดียวระหว่างคุณลักษณะนั้นๆ กับข้อมูลเอาต์พุตเท่านั้น เช่น วิธี Chi-Square เป็นต้น 2) Wrapper เป็นวิธีที่ถูกพัฒนามาเพื่อแก้ไขวิธี Filter ซึ่งคุณลักษณะทั้งหมดจะถูกจัดให้อยู่ในรูปของเซตคุณลักษณะ หลังจากนั้นดำเนินการค้นหาเซตของคุณลักษณะที่เหมาะสมด้วยการประเมินด้วยฟังก์ชันความเหมาะสม เช่น วิธี Sequential Forward Selection: SFS, Sequential Backward Selection: SBS เป็นต้น 3) Embedded หรือแบบฝังตัว คือ การใช้หลักการของ Wrapper ในการเลือกคุณลักษณะ และใช้วิธี Filter เลือกจำนวนคุณลักษณะ ซึ่งช่วยลดความซับซ้อนในการคำนวณ เนื่องจากวิธี Filter เลือกจำนวนคุณสมบัติที่มีประสิทธิภาพอย่างรวดเร็ว มีงานวิจัยเกี่ยวกับการเลือกคุณลักษณะแบบฝังตัวโดยใช้ความน่าจะเป็นตามแบบจำลองการเพิ่ม

ประสิทธิภาพ (Saito *et al.*, 2018) ทำการเปรียบเทียบเลือกคุณลักษะของกับวิธี Filter และวิธี Wrapper จากชุดข้อมูล 3 ชุดที่มีจำนวนคุณลักษณะไม่เท่ากัน ผลการวิจัยพบว่าวิธี Embedded มีประสิทธิภาพดีที่สุด ดังนั้นในงานวิจัยนี้จึงได้ทำการเปรียบเทียบการเลือกคุณลักษณะทั้ง 3 ประเภท จำนวน 4 วิธี ได้แก่ 1) Chi-Square 2) SFS 3) SBS และ 4) Embedded ร่วมกับอัลกอริทึม 3 อัลกอริทึม ได้แก่ 1) AdaBoost 2) GBM และ 3) XGBoost ผลการเปรียบเทียบการประเมินประสิทธิภาพสำหรับการเลือกคุณลักษณะ พบว่าวิธี Embedded มีประสิทธิภาพสูงสุด แสดงการเปรียบเทียบดัง Figure 9 และสามารถเลือกคุณลักษณะได้จำนวน 9 คุณลักษณะจากการพิจารณาค่าผลรวม (Total) ที่มีค่ามากกว่า 0 แสดงดัง Figure 10 ดังนั้นจึงนำคุณลักษณะที่เลือกนำไปเรียนรู้กับแบบจำลอง

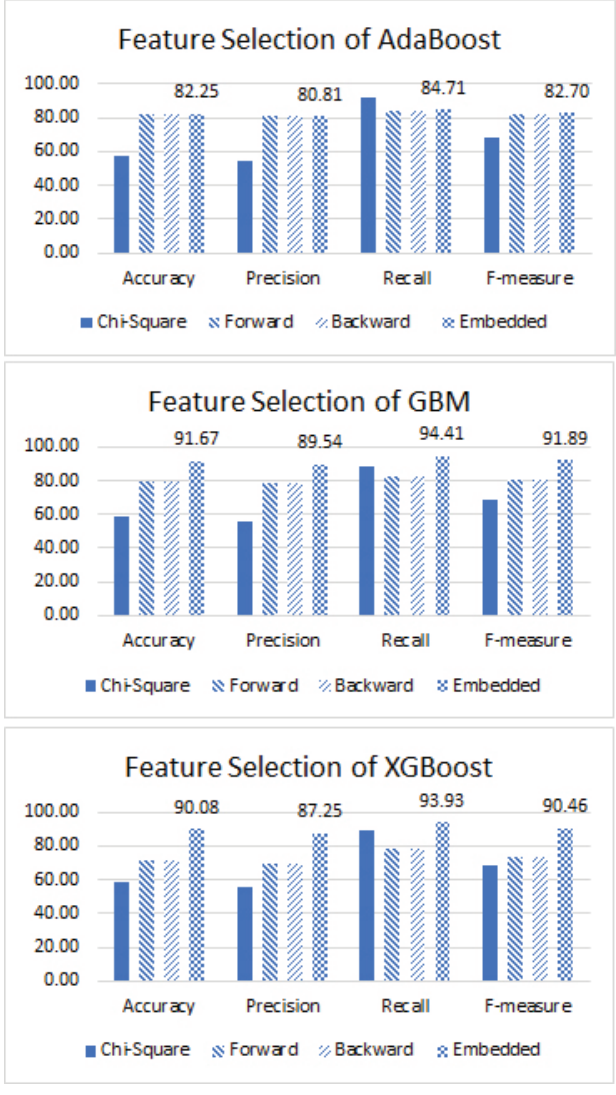

Figure 9 Performance comparison results for feature selection

3. สร้างแบบจำลอง โดยใช้อัลกอริทึม 3 อัลกอริทึม ได้แก่ 1) AdaBoost 2) GBM และ 3) XGBoostและใช้วิธีการแบ่งข้อมูลเพื่อเรียนรู้และทดสอบด้วยวิธีตรวจสอบแบบไขว้ (K-Fold Cross Validation) (Daniel Berrar, 2018) เป็นวิธีการสุ่มตัวอย่างข้อมูลที่แบ่งชุดข้อมูลการเรียนรู้จะถูกแบ่งออกเป็นส่วนย่อยจำนวน K ที่มีขนาดเท่ากันโดยประมาณ จากนั้นข้อมูลหนึ่งส่วนจะใช้เป็นตัวทดสอบประสิทธิภาพของแบบจำลอง ทำวนไปเช่นนี้จนครบจำนวนที่แบ่งไว้ ซึ่งในงานวิจัยนี้ได้นำวิธีการดังกล่าวมาประยุกต์ใช้ในการแบ่งชุดข้อมูล โดยทำการเปรียบเทียบค่า K-Fold จำนวน 3 ค่า ได้แก่ 10 15 และ 20 เพื่อหาค่าที่เหมาะสมกับอัลกอริทึม ผลการเปรียบเทียบพบว่าค่า K-Fold ที่เหมาะสมกับอัลกอริทึม คือ 10 และเมื่อค่า K เพิ่มมากขึ้นจะทำให้ค่าความถูกต้อง ของทุกๆ อัลกอริทึมลดลง

4. ประเมินประสิทธิภาพแบบจำลอง โดยเปรียบเทียบประสิทธิภาพของแบบจำลอง เพื่อทำนายพฤติกรรมการผิดเงื่อนไขการปล่อยชั่วคราวของศาล ผลการประเมินประสิทธิภาพแบบจำลองพบว่าแบบจำลองที่ใช้อัลกอริทึม GBM มีประสิทธิภาพสูงสุด ซึ่งมีค่า Accuracy 91.27% ค่า Precision 89.28% ค่า Recall 93.86% ค่า F-measure 91.51% แสดงดัง Figure 11

	Feature	AdaBoostt	GBM	XGBoost	Total
1	occupation	True	True	True	3
2	age	True	True	True	3
3	guarantee_price	True	True	False	2
4	land	True	False	True	2
5	insurance	True	False	True	2
6	plaint	True	True	False	2
7	locale	True	False	False	1
8	gender	False	False	True	1
9	unsecured	False	False	True	1
10	class_case	False	False	False	0
11	defendant_no	False	False	False	0
12	plaintiff	False	False	False	0
13	EM	False	False	False	0
14	saving_lottery	False	False	False	0
15	passbook	False	False	False	0
16	person	False	False	False	0
17	cash	False	False	False	0
18	swear	False	False	False	0

Figure 10 Feature selection results

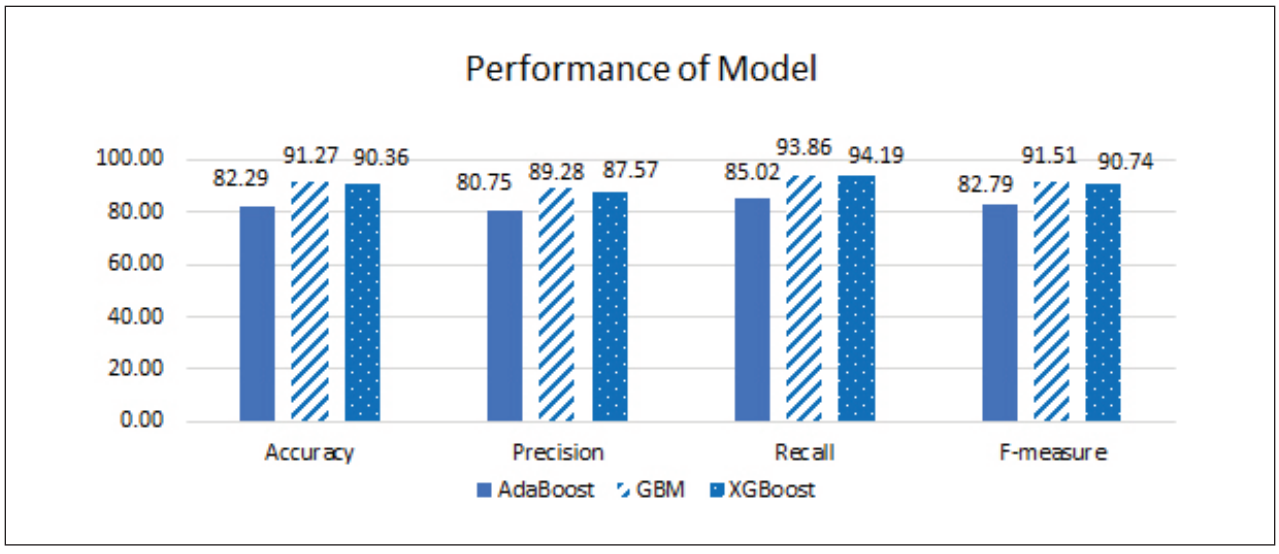

Figure 11 Model performance evaluation results

5. การปรับพารามิเตอร์เพื่อหาค่าที่เหมาะสมที่สุดสำหรับอัลกอริทึมทั้ง 3 อัลกอริทึม ด้วยวิธี GridSearch (Elgeldawi *et al.*, 2021) เป็นการค้นหาแบบกริดจะฝึกอัลกอริทึมการเรียนรู้ของเครื่องกับชุดค่าไฮเปอร์พารามิเตอร์ทั้งหมด โดยใช้เทคนิค cross-validation เทคนิคนี้ช่วยให้มั่นใจว่าแบบจำลองได้รับการฝึกจากรูปแบบส่วนใหญ่ของชุดข้อมูล ด้วยการสร้างกริดทั้งหมดที่เป็นไปได้จะทำการคำนวณค่าของแต่ละแบบจำลองเพื่อประเมิน แล้วเลือกแบบจำลองที่ให้ผลลัพธ์ดีที่สุด ซึ่งมีงานวิจัยเกี่ยวกับการวินิจฉัยโรคพาร์กินสันโดยใช้การเรียนรู้ของเครื่อง (หัสพล ธัมมิกรัตน์, 2563) ใช้การปรับค่าพารามิเตอร์เพื่อหาความเหมาะสมที่สุด ด้วยวิธี GridSearch ผลทำให้ได้ค่าความถูกต้องร้อยละ 88.78 ซึ่งงานวิจัยนี้จึงได้นำวิธี GridSearch มาใช้ปรับค่าพารามิเตอร์ที่เหมาะสม แสดงผลดัง Table 1

Table 1 The appropriate parameters of the algorithm

Tuning parameters	Search range	Parameters
learning_rate	0.1,0.5, 1	0.5
max_depth	3-5	4
n_estimators	100 - 500	200
subsample	0.1,0.5, 1	1

เมื่อทำการปรับค่าพารามิเตอร์ที่เหมาะสมและวัดประสิทธิภาพของแบบจำลองแล้ว พบว่าแบบจำลองมีประสิทธิภาพสูงขึ้น ซึ่งมีค่า Accuracy 97.44% ค่า Precision 96.37% ค่า Recall 98.39% ค่า F-measure 97.46% แสดงผลการประเมินประสิทธิภาพ แสดงดัง Figure 12

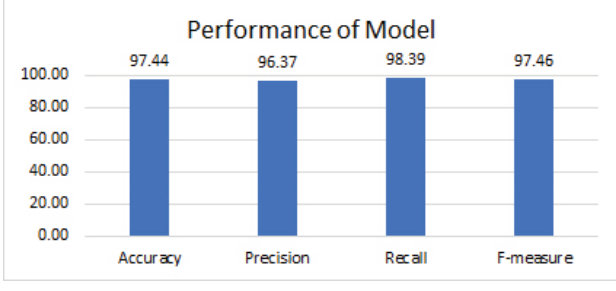

Figure 12 The results of the model performance evaluation after adjusting the appropriate parameters

6. การนำแบบจำลองไปใช้งาน โดยทำการทำนายกับชุดข้อมูลใหม่ที่ไม่สมดุล ซึ่งมีอัตราส่วน 1: 24.26 จำนวน 252 ระเบียน (record) โดยทำการประเมินประสิทธิภาพแบบจำลอง พบว่ามีค่าความถูกต้อง (Accuracy) เท่ากับ 89.68% และประเมินประสิทธิภาพความถูกต้องสำหรับการทำนายผลของแบบจำลองด้วยวิธี Receiver Operating Characteristics: ROC หรือ Area Under the Curve: AUC (สลิลย เศษเพ็ง และคณะ, 2563) ซึ่งเป็นวิธีที่ใช้ในการประเมินความถูกต้องสำหรับการทำนายผลของแบบจำลอง โดยตัวแปรที่ใช้บอกความถูกต้องของแบบจำลองได้แก่ ความถูกต้อง (Accuracy) และพื้นที่ใต้เส้นโค้ง ROC หรือ AUC เกิดจากจุดคู่อันดับระหว่าง True Positive Rate: TPR และ False Positive Rate: FPR สร้างเป็นเส้นโค้ง และคำนวณพื้นที่ใต้เส้นโค้งนั้น มีค่าอยู่ระหว่าง 0-1 ยิ่งเข้าใกล้ 1 แบบจำลองในภาพรวมสามารถทำนายได้ดีมาก ซึ่งผลการประเมินประสิทธิภาพพบว่ามีค่า AUC เท่ากับ 0.80 แสดงดัง Figure 13

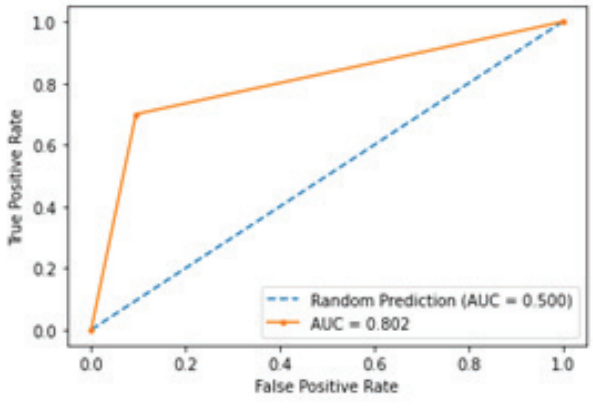

Figure 13 The result of evaluating the ground under the curve

ผลการศึกษาและอธิปรายผล

จากผลการวิจัยพบว่า การแก้ไขปัญหาความไม่สมดุลของข้อมูลด้วยวิธีการสุ่มตัวอย่างสังเคราะห์ที่ปรับเปลี่ยนได้ ADASYN และสร้างแบบจำลองด้วยอัลกอริทึม Gradient Boosting ใช้วิธีการตรวจสอบแบบไขว้ (K-Fold Cross Validation) โดยมีค่า K-Fold เท่ากับ 10 ในการแบ่งเป็นชุดข้อมูลเพื่อเรียนรู้และชุดข้อมูลทดสอบ ซึ่งได้ค่าประสิทธิภาพที่ดีที่สุดสำหรับการทำนายพฤติกรรมการผิดเงื่อนไขการปล่อยชั่วคราวของศาล ซึ่งได้ผลคล้ายกับงานวิจัยของพุทธิพร ธนธรรมเมธี และเยาวเรศ ศิริสถิตกุล และงานวิจัยของ He และคณะ ผลการแก้ไขปัญหาความไม่สมดุลของข้อมูลด้วยวิธี ADASYN ได้ค่าประสิทธิภาพที่ดีที่สุดเช่นกัน เนื่องจาก ADASYN สร้างตัวอย่างสังเคราะห์ (Synthetic Sampling) ที่ไม่จำเป็นต้องพิจารณาข้อมูลทุกตัวที่อยู่ในกลุ่มน้อย โดยจะใช้ค่าการแจกแจงแบบถ่วงน้ำหนัก (weight distribution) ของข้อมูลตัวอย่างในกลุ่มน้อย และเป็นวิธีที่ปรับปรุงการทำงานของ SMOTE ให้ดีขึ้น ซึ่งช่วยลดความเอนเอียงที่เกิดจากความไม่สมดุลของข้อมูล และทำให้มีการปรับขอบเขตของการตัดสินใจในการจำแนกกลุ่มดีขึ้น

ผลการวิจัยนี้ได้เพิ่มประสิทธิภาพให้แบบจำลองด้วยการใช้วิธีการเลือกคุณลักษณะ (feature selection) แบบฝังตัว (embedded approach) ทำการประเมินผลการเลือกคุณลักษณะที่มีผลต่อประสิทธิภาพของแบบจำลองได้ จำนวน 9 คุณลักษณะ เมื่อพิจารณาชุดข้อมูลสำหรับคุณลักษณะที่ไม่ถูกเลือกไปเรียนรู้กับแบบจำลอง พบว่าข้อมูลในคุณลักษณะดังกล่าวมีความไม่สมดุล เช่น คุณลักษณะ EM การติดอุปกรณ์อิเล็กทรอนิกส์ติดตามตัวซึ่งข้อมูลผู้ที่ติดอุปกรณ์ EM มีจำนวน 7 ระเบียน ต่อผู้ไม่ติดอุปกรณ์ EM 2,570 ระเบียน คิดเป็นอัตราส่วนความไม่สมดุลเท่ากับ 1:367.14 ซึ่งมีผลต่อประสิทธิภาพของแบบจำลอง เป็นต้น และทำการปรับค่าพารามิเตอร์ที่เหมาะสมที่สุด (parameter optimal) แบบอัตโนมัติด้วยวิธี GridSearch ซึ่งมีผลต่อประสิทธิภาพแบบจำลองมีประสิทธิภาพสูงขึ้น และทดสอบแบบจำลองกับข้อมูลชุดใหม่ พบว่ามีประสิทธิภาพความถูกต้อง 89.68% และค่า AUC เท่ากับ 0.80 ซึ่งสามารถนำแบบจำลองไปประยุกต์ใช้ในอนาคตได้ หากจะนำงานวิจัยไปพัฒนาให้มีประสิทธิภาพมากขึ้น ควรเพิ่มข้อมูลคุณลักษณะ (feature) ที่เป็นปัจจัยหรือมีผลต่อการผิดเงื่อนไขการปล่อยชั่วคราวของศาล ดังนั้น ควรรวบรวมหรือจัดเก็บข้อมูลด้านอื่นๆ เพิ่มเติม และควรปรับเปลี่ยนอัลกอริทึมที่เหมาะสมกับชุดข้อมูลแบบจำแนกประเภทเพื่อให้ได้ประสิทธิภาพที่ดีที่สุด และขอเสนอแนะใช้การเรียนรู้เชิงลึก (Deep Learning: DL) สำหรับการสร้างแบบจำลองนี้

กิตติกรรมประกาศ

ขอขอบคุณ สำนักงานศาลยุติธรรมที่สนับสนุนทุนการศึกษา และขอขอบคุณศาลจังหวัดพะเยาที่ได้อนุญาตให้ใช้ชุดข้อมูลจากโปรแกรมศาลชั้นต้นสำหรับนำมาศึกษาในงานวิจัยนี้

เอกสารอ้างอิง

กองการต่างประเทศ สำนักงานศาลยุติธรรม. (2555). *การขอปล่อยชั่วคราวต่อศาล*. กองการต่างประเทศ สำนักงานศาลยุติธรรม

กิตติภพ แซ่เตีย และ จิรภัทร์ หยกรัตนศักดิ์. (2564). *การจัดการข้อมูลไม่สมดุลของการทำกลยุทธ์เสนอขายประกันต่อยอดสำหรับผู้ถือบัตรเครดิต*. ใน: *เอกสารการประชุมวิชาการระดับชาติ ครั้งที่ 13 มหาวิทยาลัยราชภัฏนครปฐม* (หน้า 514-523). มหาวิทยาลัยราชภัฏนครปฐม.

พุทธิพร ธนธรรมเมธี และเยาวเรศ ศิริสถิตย์กุล. (2561). เทคนิคการจำแนกข้อมูลที่พัฒนาสำหรับชุดข้อมูลที่ไม่สมดุลของภาวะข้อเข่าเสื่อมในผู้สูงอายุ. *วารสารวิทยาศาสตร์และเทคโนโลยี, 27*(6), 1164-1178.

วิชญ์วิสิฐ เกสรสิทธิ์, วิชิต หล่อจีระชุณห์กุล และจิราวัลย์ จิตรถเวช. (2561). การแก้ปัญหาข้อมูลไม่สมดุลของข้อมูลสาหรับการจาแนกผู้ป่วยโรคเบาหวาน. *KKU Research Journal (Graduate Studies), 18*(3), 11-21.

สลิลยา เศษเพ็ง, เทพไท ไชยทอง และ สุทธิศักดิ์ศรลัมพ์. (2563). *การประเมินความแม่นยำของแบบจำลองปริมาณน้ำฝนสะสมวิกฤติ (AP-Model) ในการคาดการณ์พื้นที่ระดับความอ่อนไหวต่อการเกิดดินถล่มล่วงหน้า.* การประชุมวิชาการวิศวกรรมโยธาแห่งชาติ ครั้งที่ 25, ชลบุรี ; 2563.

สำนักงานพัฒนารัฐบาลดิจิทัล (องค์การมหาชน). (2563). *กรอบการทำงานปัญญาประดิษฐ์ภาครัฐ*. https://www.dga.or.th/document-sharing/dga-e-book/annual-ai/47112/.

หัสพล ธัมมิกรัตน์. (2563). *การวินิจฉัยโรคพาร์กินสันโดยใช้การเรียนรู้ของเครื่อง*. จุฬาลงกรณ์มหาวิทยาลัย.

Berrar, D. (2018). Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology*,1, Elsevier.

Chawla, Nitesh V., Bowyer, Kevin W., Hall, Lawrance O., & Kegelmeyer, W., Philip. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321-357.

Chen, Z., Zhou, L. & Yu, W. (2021). ADASYN-random forest based intrusion detection model. *SPML 2021: 2021 4th International Conference on Signal Processing and Machine Learning* (pp. 152-159), United States.

Chen, T. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference*.

Elgeldawi, E., Sayed, A., Galal, A. & Zaki, A. (2021). Hyperparameter tuning for machine learning algorithms used for Arabic sentiment analysis. *Informatics, 8*, 79.

Gheyas, I. & Smith, L. (2010). Feature subset selectionin large dimensionality domains. *Pattern Recognition, 43*, 5-13.

Guan, J., Jiang, X & Mao, B. (2021). A method for class-imbalance learning in android malware detection. *Electronics 2021, 10*, 3124.

Han, H., Wang, W. & Mao, B. (2005). *Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning*. In: *International Conference on Intelligent Computing Hefei* (pp. 878-887). China.

He, H., B., Yang, G. E. & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *IEEE World Congress on Computational Intelligence: 2008 IEEE International Joint Conference on Neural Networks* (pp. 1322-1328). Hong Kong.

Jiang, Z., Pan, T., Zhang, C. & Yang, J. (2021). A new oversampling method based on the classification contribution degree. *Symmetry, 13*(2), 194.

Krawczyk. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence, 5*, 221-232.

Kulkarni, A., Feras, A., Batarseh & Chong, D. (2020). Foundations of data imbalance and solutions for a data democracy. *Data Democracy At the Nexus of Artificial Intelligence, Software Development, and Knowledge Engineering*, 83-106.

Minh, H. (2018, October 11). *How to Handle Imbalanced Data in Classification Problems*. https://medium.com/@nminh.hoang1023/handling-imbalanced-data-in-classification-problems-7de598c1059f.

Natekin, A. & Knoll, A. (2013). Gradient Boosting Machines. A Tutorial. *Frontiers in Neurorobotics,* 7, 21.

Saito, S., Shirakawa, S & Akimoto, Y. (2018). Embedded feature selection using probabilistic model-based. *Optimization. the Genetic and Evolutionary Computation Conference 2018 Companion* (pp. 1922-1925). Japan.

Schapire, R.E. & Freund, Y. (2012). *Boosting: foundations and algorithms*. The MIT Press Cambridge.

Wu, P. & Zhao, H. (2011). Some analysis and research of the AdaBoost algorithm. *Communications in Computer and Information Science, 134*, 1-5.

Zhua, R., Guob, Y & Xuec, J.H. (2020). Adjusting the imbalance ratio by the dimensionality of imbalanced data. *Pattern Recognition Letters, 133*, 217-223.