การทำนายการระบาดของไข้เลือดออกโดยใช้ระบบสารสนเทศภูมิศาสตร์ร่วมกับการเรียนรู้ของเครื่อง: กรณีศึกษาอำเภอเมืองพะเยา จังหวัดพะเยา

Dengue outbreak prediction using geographic information system and machine learning: A case study of Mueang Phayao District, Phayao Province

เสถียร หันตา[1*], พันธิตรา ลือชา[1] และ จิราพัชร บุญสูง[1]

Sathien Hunta[1*], Phanthitra Luecha[1] and Jirapad Boonsoong[1]

Received: 9 September 2024 ; **Revised**: 4 October 2024 ; **Accepted**: 6 November 2024

บทคัดย่อ

ไข้เลือดออกเป็นโรคติดต่อที่มียุงลายเป็นพาหะ และเป็นปัญหาสาธารณสุขในหลายประเทศทั่วโลก งานวิจัยนี้จึงมีวัตถุประสงค์เพื่อศึกษาปัจจัยที่มีผลต่อการระบาดของโรคไข้เลือดออก และสร้างแบบจำลองการพยากรณ์โรคด้วยเทคนิคการเรียนรู้ของเครื่อง ข้อมูลที่ใช้ในการศึกษารวบรวมจากสถานีอุตุนิยมวิทยาพะเยา และจำนวนผู้ป่วยจากโรงพยาบาลพะเยา ระหว่างปี 2561 ถึง 2565 โดยแยกชุดข้อมูลตามปัจจัยต่าง ๆ ได้แก่ กลุ่มเพศ กลุ่มอายุ จำนวนประชากร พื้นที่ป่าไม้ ระยะห่างระหว่างพื้นที่การระบาดกับแหล่งน้ำ ระยะห่างระหว่างพื้นที่การเกิดการระบาดซ้ำ อุณหภูมิเฉลี่ย ความชื้นสัมพัทธ์เฉลี่ย ปริมาณน้ำฝนเฉลี่ย และจำนวนวันที่ฝนตก อีกทั้งได้มีการประยุกต์ใช้ระบบสารสนเทศภูมิศาสตร์เข้ามาช่วยในการวิเคราะห์หาปัจจัยที่เกี่ยวข้อง

ข้อมูลทั้งหมดถูกนำมาสร้างเป็น dataset แบบรายเดือน และรายสัปดาห์ โดยการสร้างแบบจำลองจะแบ่งออกเป็น 2 ประเภท คือ ประเภท Regression ได้แก่ Support Vector Machines, Generalized Linear Model และ Artificial Neural Network และ ประเภท Classification ได้แก่ Decision Tree, Gradient Boosting Tree และ Random Forest โดยจะใช้วิธี Ensemble Method แบบ Stacking Model เพื่อรวมหลายอัลกอริทึมการเรียนรู้ของเครื่องเข้าด้วยกันเพื่อปรับปรุงประสิทธิภาพในการพยากรณ์ โดยใช้ 2 อัลกอริทึมเป็น Base Learner และใช้อัลกอริทึมที่เหลือเป็น Stacking Model Learner จากนั้นใช้ 5-fold Cross Validation เพื่อเปรียบเทียบ และวัดประสิทธิภาพของแบบจำลอง

ผลการศึกษาพบว่าประเภท Regression การใช้ Generalized Linear Model เป็น Stacking Model Learner มีประสิทธิภาพที่ดีที่สุด โดยมีค่า RMSE 1.457 ส่วนประเภท Classification แบบจำลอง Gradient Boosting Tree มีประสิทธิภาพสูงสุด โดยมีค่า Accuracy 99.12% การศึกษานี้ชี้ให้เห็นถึงความสำคัญของการใช้เทคนิคการเรียนรู้ของเครื่องในการพยากรณ์ ซึ่งอาจนำไปใช้เป็นเครื่องมือในการเฝ้าระวัง และควบคุมโรคได้อย่างมีประสิทธิภาพในอนาคต

คำสำคัญ: ไข้เลือดออก, การเรียนรู้ของเครื่อง, แบบจำลองการพยากรณ์, ระบบสารสนเทศภูมิศาสตร์, การเรียนรู้แบบผสมผสาน

Abstract

Dengue fever is a mosquito-borne infectious disease and is a public health problem in many countries around the world. This research aims to study the factors affecting the spread of dengue fever and to create prediction models using machine learning techniques. The data were collected from the Phayao Meteorological Station and Phayao Hospital between 2018 and 2022. The geographic information system was used to analyze the related factors. The data preparation was performed by separating the datasets according to various factors, including gender, age group, population, forest area, distance between outbreak areas and water sources, distance between re-outbreak areas, average temperature, average relative humidity, average rainfall, and number of rainy days.

---

[1] *คณะเทคโนโลยีสารสนเทศและการสื่อสาร มหาวิทยาลัยพะเยา จังหวัดพะเยา 56000*

[1] *School of Information and Communication Technology, University of Phayao, Phayao, 56000*

** Corresponding author:sathien.hu@up.ac.th*

All data were generated as monthly and weekly datasets. The prediction models are divided into regression and classification types. The regression includes support vector machines, generalized linear models, and artificial neural networks. The classification includes decision trees, gradient boosting trees, and random forests. The ensemble method of stacking models is used to combine multiple machine learning algorithms to improve the performance by using 2 algorithms as base learners and using another algorithm as the stacking model learner. Performance evaluation was performed through five-fold cross-validation and the different datasets and learning methods were compared.

The results found that the regression type using the generalized linear model as the stacking model learner gave the best performance with an RMSE value of 1.457, while the classification using the gradient boosting tree gave the highest performance with an accuracy of 99.12%. This study highlights the importance of using machine learning techniques in forecasting dengue fever outbreaks, which can be used as an effective tool for disease surveillance and control in the future.

Keywords: Dengue fever, machine learning, prediction model, geographic information system, ensemble learning

บทนำ

โรคไข้เลือดออกเป็นโรคร้ายแรงที่ติดต่อได้จากเชื้อไวรัสเดงกี ซึ่งถูกแพร่ผ่านทางยุงลายบ้าน ไข้เลือดออก (Dengue fever) มีสาเหตุมาจากเชื้อไวรัสเดงกีที่ยุงลายที่เป็นพาหะนำมาเมื่อมันกัดคนที่ติดเชื้อ ไวรัสเดงกีมี 4 สายพันธุ์ ได้แก่ DENV – 1, DENV – 2, DENV – 3 และ DENV – 4 ที่ทำให้เกิดไข้เลือดออกได้ โรคไข้เลือดออกถ้าหากเคยติดเชื้อซ้ำแล้วสามารถติดเชื้อซ้ำได้อีก ในการติดเชื้อซ้ำครั้งถัดไป อาจเพิ่มความเสี่ยงในการเป็นโรคไข้เลือดออกที่รุนแรงขึ้นถึงร้อยละ 80 - 90 เนื่องจากมีไวรัสถึง 4 สายพันธุ์ และผู้ป่วยโรคไข้เลือดออกที่เคยได้รับเชื้อไวรัสสายพันธุ์ใด ๆ จะมีภูมิคุ้มกันเฉพาะสายพันธุ์นั้นไปตลอดชีวิต และจะมีภูมิคุ้มกันต่อสายพันธุ์อื่น 6 - 12 เดือนเท่านั้น (โรงพยาบาลพญาไท, ม.ป.ป.) ผู้ที่ติดเชื้อไวรัสจะมีอาการหลัก ๆ ได้แก่ ไข้สูง ปวดกล้ามเนื้อ ปวดหัว คลื่นไส้อาเจียน และมีจ้ำเลือดหรือจุดเลือดบนผิวหนัง ในกรณีอาการรุนแรงสามารถทำให้เกิดภาวะไหลเวียนโลหิตล้มเหลว ช็อกหรือเสียชีวิตได้ ดังนั้น หากมีความเสี่ยงที่จะติดเชื้อไวรัสไข้เลือดออก ควรพบแพทย์เพื่อรับการตรวจ และรักษาโดยทันที (กรมควบคุมโรค, 2563)โดยที่การแพร่ระบาดของโรคไข้เลือดออกเป็นปัญหาสาธารณสุขที่สำคัญในหลายพื้นที่ของประเทศไทย โดยเฉพาะจังหวัดพะเยา ซึ่งมีการระบาดของโรคไข้เลือดออกเป็นประจำ โดยพบผู้ป่วยโรคไข้เลือดออกเป็นจำนวนมากในพื้นที่อำเภอเมืองพะเยาของทุก ๆ ปี ในประเทศไทยเริ่มมีการรายงานการติดเชื้อโรคไข้เลือดออกครั้งแรกในปี พ.ศ. 2492 โดยเกิดการระบาดครั้งแรกในปี พ.ศ. 2501 ระบาดใหญ่ที่สุดในปี พ.ศ. 2530 ซึ่งพบผู้ป่วยสูงสุดถึงกว่า 170,000 ราย และจำนวนผู้เสียชีวิตมากกว่า 1,000 ราย (กรมควบคุมโรค, 2560) ส่วนในพื้นที่จังหวัดพะเยา มีการรายงานสถานการณ์โรคไข้เลือดออก ตั้งแต่วันที่ 1 มกราคม – 8 กันยายน ปี พ.ศ. 2566 พบผู้ป่วยในทุกอำเภอ และมียอดผู้ป่วยโรคไข้เลือดออกสะสม 951 ราย ผู้เสียชีวิตจำนวน 1 ราย (สำนักงานสาธารณสุขจังหวัดพะเยา, 2566) การระบาดเหล่านี้ส่งผลกระทบต่อสุขภาพของประชาชน ซึ่งก่อให้เกิดภาระทางการแพทย์ และเศรษฐกิจอย่างมาก

ระบบสารสนเทศภูมิศาสตร์ (Geographic Information System: GIS) คือระบบที่รวบรวม จัดการ หรือวิเคราะห์ข้อมูลทางภูมิศาสตร์ เช่น ภาพถ่ายจากดาวเทียม ภาพแผนที่ ตัวเลข ตัวอักษร ระยะทาง และอื่น ๆ โดยวิเคราะห์ผ่านทางโปรแกรมคอมพิวเตอร์ สามารถนำไประยุกต์ใช้งานได้หลายด้านโดยเฉพาะการจัดการสิ่งแวดล้อม การคำนวณหาพื้นที่บริเวณที่จะใช้งานจากภาพแผนที่ การวัดระยะทางในการใช้งาน (ศูนย์วิจัยภูมิสารสนเทศเพื่อประเทศไทย, ม.ป.ป)

กระบวนการของ Data Mining ประกอบด้วยขั้นตอนทั้งหมด 5 ขั้นตอน คือ 1) การทำความเข้าใจข้อมูล เริ่มต้นด้วยการวิเคราะห์ข้อมูลที่ได้จากการสำรวจเพื่อสร้างความคุ้นเคยกับข้อมูล และค้นหาความเข้าใจเชิงลึกเบื้องต้น ในขั้นตอนนี้จะใช้เทคโนโลยีสารสนเทศภูมิศาสตร์ (GIS) เพื่อสร้างฐานข้อมูลที่เกี่ยวข้องกับปัจจัยต่าง ๆ และทำความเข้าใจปัญหาผลกระทบ หรืออุปสรรคที่เกี่ยวข้องกับโครงการได้ 2) การเตรียมข้อมูลโดยทำการแปลงข้อมูลดิบให้อยู่ในรูปแบบที่เหมาะสมสำหรับการวิเคราะห์ โดยในขั้นตอนนี้ควรพิจารณาแหล่งที่มาของข้อมูล ความแม่นยำ ความน่าเชื่อถือของข้อมูล รวมถึงปริมาณ และรายละเอียดของข้อมูลที่เหมาะสมสำหรับการทำเหมืองข้อมูล 3) การสร้างแบบจำลองการพยากรณ์

โดยการเลือกเทคนิคหรืออัลกอริทึมที่เหมาะสมสำหรับการสร้างแบบจำลองการพยากรณ์ โดยต้องตอบสนองต่อวัตถุประสงค์ของการวิเคราะห์ และการพยากรณ์อย่างชัดเจน 4) การประเมินผล ซึ่งเป็นการประเมินผลลัพธ์ของแบบจำลองที่สร้างขึ้น โดยใช้ข้อมูลทดสอบเพื่อตรวจสอบความถูกต้องความแม่นยำ และ 5) การนำไปใช้ เป็นการนำแบบจำลองที่ผ่านการประเมิน หรือปรับแต่งแล้วไปใช้งานจริง เพื่อให้เกิดประโยชน์ในทางปฏิบัติ (สายชล สินสมบูรณ์ทอง, 2560)

การเรียนรู้ของเครื่อง (Machine Learning) เป็นศาสตร์ย่อยแขนงหนึ่งของ Artificial Intelligence (AI) ที่ช่วยให้คอมพิวเตอร์สามารถทำงานที่ซับซ้อนได้อย่างมีประสิทธิภาพและอัตโนมัติ ซึ่งจะแตกต่างกันตรงในส่วนของโปรแกรมที่คำนวณประมวลผล การตั้งเงื่อนไข และแบบจำลองจะถูกสร้างโดยคอมพิวเตอร์ โดยจะนำชุดข้อมูล (Dataset) ทั้ง Data และ Output เข้าไปให้คอมพิวเตอร์เรียนรู้เพื่อสร้างแบบจำลอง ช่วยให้คอมพิวเตอร์สามารถทำงานอย่างมีประสิทธิภาพมากขึ้นในหลากหลายด้าน เช่น การรู้จำภาพ การวิเคราะห์ภาษา การพยากรณ์ผลในอนาคต การแนะนำสินค้า และอื่น ๆ (กอบเกียรติ สระอุบล, 2563)

ผู้วิจัยจึงใช้ระบบสารสนเทศภูมิศาสตร์เข้ามาช่วยในการหาปัจจัยด้านสารสนเทศภูมิศาสตร์ และนำปัจจัยที่เกี่ยวข้องมาพยากรณ์โรคไข้เลือดออกโดยใช้เทคนิคการเรียนรู้ของเครื่อง เพื่อให้มีแนวทางการป้องกันล่วงหน้าที่มีประสิทธิภาพและลดอัตราการเกิดโรคไข้เลือดออกในอำเภอเมืองพะเยา จังหวัดพะเยา

ขอบเขตพื้นที่ศึกษา

พื้นที่การศึกษาในการวิจัยครั้งนี้คืออำเภอเมืองพะเยา จังหวัดพะเยา ครอบคลุมพื้นที่ตำบลทั้งหมด 15 ตำบล ได้แก่ ตำบลเวียง ตำบลแม่ต๋ำ ตำบลแม่นาเรือ ตำบลบ้านตุ่น ตำบลบ้านต๋ำ ตำบลบ้านต๋อม ตำบลแม่ปืม ตำบลแม่กา ตำบลจำป่าหวาย ตำบลบ้านใหม่ ตำบลท่าวังทอง ตำบลแม่ใส ตำบลบ้านสาง ตำบลท่าจำปี และตำบลสันป่าม่วง ดังแสดงใน Figure 1 พื้นที่ดังกล่าวมีพื้นที่ 1055.915 ตารางกิโลเมตร มีประชากรอาศัยอยู่ทั้งสิ้น 114,490 คน

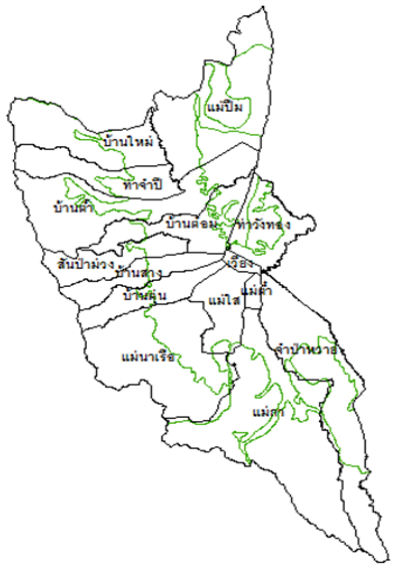


Figure 1 Study area boundaries

วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาปัจจัยที่มีผลต่อการระบาดของโรคไข้เลือดออก
2. เพื่อสร้างแบบจำลองในการพยากรณ์โรคไข้เลือดออก อำเภอเมืองพะเยา จังหวัดพะเยา
3. เพื่อทำนายจำนวนผู้ป่วยโรคไข้เลือดออก ซึ่งจะนำไปใช้ในการวางแผนการเฝ้าระวัง ป้องกัน และควบคุมโรคไข้เลือดออกในอนาคตได้อย่างมีประสิทธิภาพ

วิธีดำเนินการวิจัย

วิธีการดำเนินการวิจัยแบ่งออกเป็น 3 ขั้นตอน คือ 1) การเก็บรวมรวมข้อมูล 2) การวิเคราะห์ข้อมูล 3) การสร้างแบบจำลองการพยากรณ์ และวัดประสิทธิภาพ

1) การเก็บรวบรวมข้อมูล

ผู้วิจัยได้รวบรวมข้อมูลที่เกี่ยวกับปัจจัยที่เกี่ยวข้องจาก 3 แหล่ง คือ โรงพยาบาลพะเยา ประกอบด้วย วัน/เดือน/ปี ที่เข้ารับบริการ เพศ (ชายและหญิง) อายุ (แบ่งช่วงกลุ่มอายุเป็น 5 ช่วง โดยช่วงที่ 1 อายุ 0 – 15 ปี ช่วงที่ 2 อายุ 16 – 30 ปี ช่วงที่ 3 อายุ 31 – 45 ปี ช่วงที่ 4 อายุ 46 – 60 ปี และช่วงที่ 5 อายุ 61 ปีขึ้นไป) อำเภอ ตำบล หมู่ และ จำนวนผู้ป่วย ส่วนสถานีอุตุนิยมวิทยาพะเยา ประกอบด้วยอุณหภูมิสูงสุด – ต่ำสุด อุณหภูมิเฉลี่ย ความชื้นสัมพัทธ์เฉลี่ย ปริมาณน้ำฝนเฉลี่ย จำนวนวันที่ฝนตก และข้อมูลประชากรรวมจากฐานข้อมูล สำนักงานสาธารณสุขจังหวัดพะเยา โดยใช้ข้อมูลตั้งแต่ปี พ.ศ. 2561 – 2565

ข้อมูลสารสนเทศภูมิศาสตร์ ผู้วิจัยได้ใช้ระบบสารสนเทศภูมิศาสตร์ในการคำนวณหาพื้นที่ และระยะห่างของ

ปัจจัย ได้แก่ 1) พื้นที่ป่าไม้ 2) ระยะห่างระหว่างพื้นที่การระบาดกับแหล่งน้ำ 3) ระยะห่างระหว่างพื้นที่การเกิดการระบาดซ้ำ โดยแสดงแผนที่ตำแหน่งผู้ป่วยโรคไข้เลือดออกในอำเภอเมืองพะเยา จังหวัดพะเยา ปี พ.ศ. 2561 – 2565 ดังแสดงใน Figure 2 และรายละเอียดของขั้นตอนการดำเนินงาน ดังแสดงใน Figure 3

2) การเตรียมข้อมูล

สำหรับการเตรียมข้อมูลสำหรับสร้างแบบจำลองการพยากรณ์ผู้ป่วยโรคไข้เลือดออกเป็นที่มีการใช้คุณลักษณะหลากหลาย ซึ่งรวมถึงจำนวนผู้ป่วยในแต่ละตำบล กลุ่มอายุ กลุ่มเพศ จำนวนประชากร ปริมาณน้ำฝนเฉลี่ยความชื้นสัมพัทธ์เฉลี่ย จำนวนวันที่ฝนตก โดยแยกเป็นข้อมูลรายสัปดาห์/รายเดือน ย้อนหลัง 5 ปี และข้อมูลด้านสารสนเทศภูมิศาสตร์ ได้แก่ ระยะห่างระหว่างพื้นที่การระบาดกับแหล่งน้ำ ระยะห่างระหว่างพื้นที่การเกิดการระบาดซ้ำ และพื้นที่ป่าไม้ ข้อมูลที่ใช้ในการวิเคราะห์ประกอบด้วยจำนวนตัวอย่างที่มีการแบ่งข้อมูลเป็นชุดสำหรับการสอนและทดสอบ ซึ่งช่วยในการประเมินความสามารถของแบบจำลองในการพยากรณ์ นอกจากนี้ยังมีการจัดการกับข้อมูลที่ขาดหายโดยการใช้เทคนิคการเติมค่าที่ขาดหาย (Imputation) ด้วยการใช้ค่าเฉลี่ย (Mean) เพื่อรักษาความสมบูรณ์ของข้อมูล และใช้ Operator Normalization ในการปรับขนาดข้อมูล ในโปรแกรม RapidMiner (Hofmann & Klinkenberg, 2013) ซึ่งจะถูกปรับโดยอัตโนมัติ เพื่อให้แน่ใจว่าค่าของคุณลักษณะต่าง ๆ อยู่ในช่วงที่เหมาะสมสำหรับการเรียนรู้ของเครื่อง ซึ่งจะช่วยเพิ่มประสิทธิภาพของโมเดลในการพยากรณ์ค่าของผู้ป่วยได้อย่างแม่นยำยิ่งขึ้น

3) การวิเคราะห์ข้อมูล

ผู้วิจัยได้วิเคราะห์ข้อมูลโดยใช้ข้อมูลจำนวนผู้ป่วยที่ได้จากโรงพยาบาลพะเยา ข้อมูลสภาพอากาศจากสถานีอุตุนิยมวิทยาพะเยา ข้อมูลประชากรจากฐานข้อมูล สำนักงานสาธารณสุขจังหวัดพะเยา และข้อมูลที่ได้มาจากระบบสารสนเทศภูมิศาสตร์ โดยแยกข้อมูลออกเป็นชุดข้อมูล โดยทำการแยกเป็นรายสัปดาห์ รายเดือน โดยใช้ข้อมูลตั้งแต่ปี พ.ศ. 2561 - 2565 โดยรายละเอียดของปัจจัย ดังแสดงใน Table 1 และแสดงรายละเอียดของชุดข้อมูล ดังแสดงใน Table 2 เพื่อนำมาสร้างแบบจำลองการพยากรณ์ด้วยโปรแกรม RapidMiner

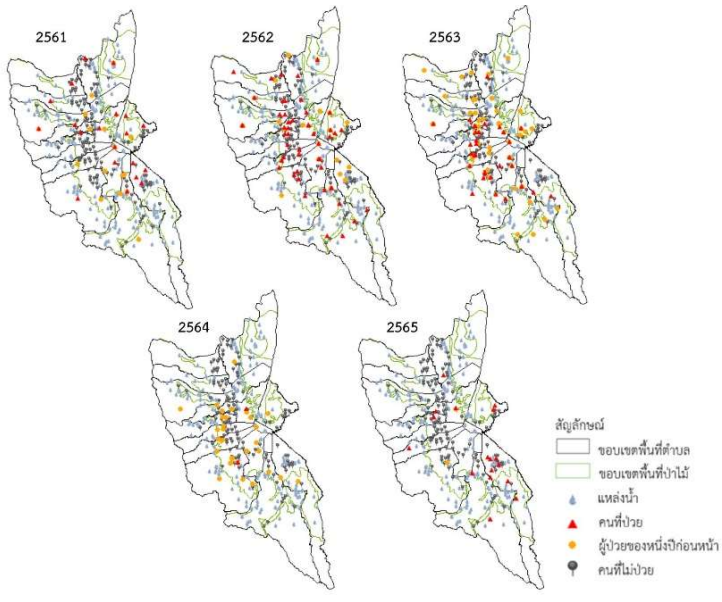


Figure 2 Map of dengue fever patient locations in Mueang Phayao District, Phayao Province, 2018 - 2022

Table 1 Factors used in the study.

Factor	Description	Data type	Source
Patient	Number of dengue fever patients	Label	Phayao Hospital
Subdistrict	Distriot	Feature	
Week	Week	Feature	
Month	Month	Feature	
Year	Year	Feature	
Age	Age group	(Feature	
Sex	Gender group	Feature	
Population	Population	Feature	Phayao Provincial Public Health Office
Precipitation	Average weekly/monthly rainfall (mm)	Feature	Phayao Meteorological Station
Humidity	Average weekly/monthly relative humidity (percent)	Feature	
Rain-day	Number of rainy days weekly/monthly	Feature	Phayao Meteorological Station
Distance - water source	Distance between outbreak are water source (meter/measure from center)	Feature	Geographic Information System
Distance - repeated outbreaks	Distance between reinfestation areas (meter/measure from center)	Feature	
Forest	Forest area (square meter)	Feature	

ปัจจัยด้านสารสนเทศภูมิศาสตร์ ในส่วนของพื้นที่ป่าไม้เมื่อพื้นที่ป่าถูกทำลาย มักจะมีการขยายตัวของพื้นที่เมืองหรือพื้นที่เกษตร ซึ่งสามารถนำไปสู่การเพิ่มขึ้นของประชากรยุงลาย และความเสี่ยงในการแพร่ระบาดของโรคไข้เลือดออก (EFD Initiative, 2024) ส่วนระยะห่างระหว่างพื้นที่การระบาดกับแหล่งน้ำ โดยทั่วไปแล้วพื้นที่ใกล้กับแหล่งน้ำมีความเสี่ยงสูงกว่า เนื่องจากยุงลายซึ่งเป็นพาหะของโรคไข้เลือดออกมักวางไข่ในน้ำขัง แหล่งน้ำที่มีการไหลน้อย และระยะห่างระหว่างพื้นที่การระบาดซ้ำมีความสำคัญ เนื่องจากผู้ที่เคยติดเชื้อไวรัสไข้เลือดออกสามารถติดเชื้อซ้ำได้ ซึ่งอาจเพิ่มความเสี่ยงในการเป็นโรคไข้เลือดออกรุนแรงขึ้น เพราะไวรัสไข้เลือดออกมีถึง 4 สายพันธุ์ ผู้ป่วยโรคไข้เลือดออกที่เคยได้รับเชื้อไวรัสสายพันธุ์ใด ๆ จะมีภูมิคุ้มกันเฉพาะสายพันธุ์นั้นไปตลอดชีวิต และจะมีภูมิคุ้มกันต่อสายพันธุ์อื่น 6 - 12 เดือน ดังนั้นผู้วิจัยจึงนำระยะห่างระหว่างพื้นที่การเกิดกับการระบาดในปีที่ผ่านมาเป็นปัจจัยหนึ่งในการสร้างแบบจำลอง

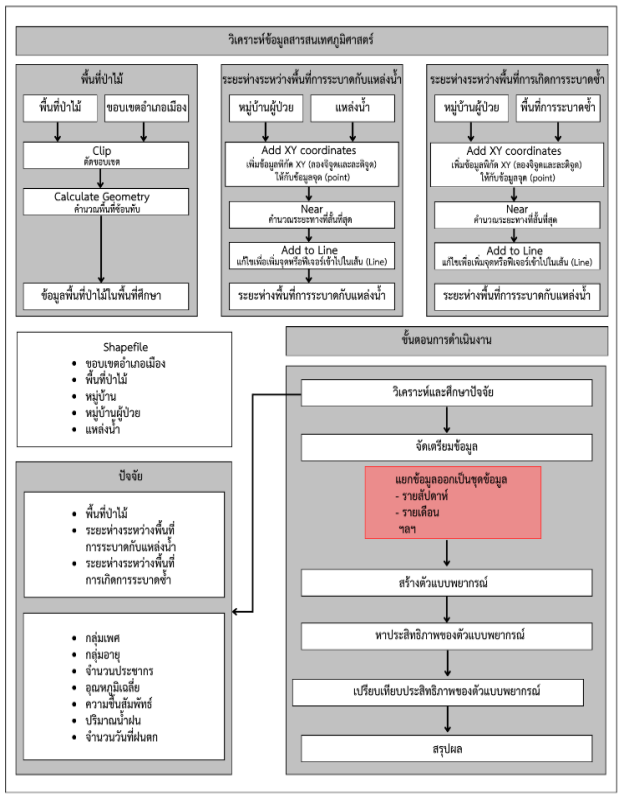


Figure 3 Procedures for implementation

4) การสร้างแบบจำลองการพยากรณ์และวัดประสิทธิภาพ

ในการดำเนินงานวิจัยได้ใช้กระบวนการของ Data mining ดังแสดงใน Figure 4 ด้วยการแปลงข้อมูลให้อยู่ในรูปแบบที่ต้องการ และสร้างแบบจำลองการพยากรณ์ เพื่อทำนายโรคไข้เลือดออก ด้วยเทคนิคการเรียนรู้ของเครื่อง โดยประเภท Regression ได้แก่ ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines) แบบจำลองเชิงเส้นทั่วไป (Generalized Linear Model) และ โครงข่ายประสาทเทียม (Artificial Neural Network) ส่วนประเภท Classification ได้แก่ ต้นไม้ตัดสินใจ (Decision Tree) อัลกอริทึมการส่งเสริมประเภทหนึ่ง (Gradient Boosting Tree) และ ป่าไม้ตัดสินใจ (Random Forest) โดยใช้วิธี 5-fold Cross Validation ซึ่งแบ่งข้อมูลออกเป็น 5 ส่วนเท่ากัน ในแต่ละรอบจะใช้ 4 ส่วนในการฝึกแบบจำลอง และอีก 1 ส่วนสำหรับทดสอบ ทำซ้ำ 5 รอบ โดยสลับส่วนที่ใช้ทดสอบในแต่ละครั้ง (Kohavi, 1995) เพื่อทำการเปรียบเทียบ และวัดประสิทธิภาพของแบบจำลองการพยากรณ์ จากนั้นนำแบบจำลองที่มีประสิทธิภาพสูงสุดมาฝึกและทดสอบอีกครั้ง โดยใช้วิธีการ Ensemble Method โดยใช้เทคนิคย่อย Stacking Model เป็นแนวทางที่สำคัญในการพัฒนาประสิทธิภาพของการพยากรณ์และการวิเคราะห์ข้อมูล โดยมีหลักการรวมแบบจำลองหลายแบบที่ถูกสร้างจากอัลกอริทึมต่าง ๆ เพื่อเพิ่มความแม่นยำของผลลัพธ์ การทำงานของ Stacking มีขั้นตอนหลักสองส่วน ได้แก่ Base Learners และ Meta Learner โดยในขั้นตอนแรก Base Learners จะทำการฝึกอบรมแบบจำลองโดยใช้อัลกอริทึมที่หลากหลายและมีคุณสมบัติต่างกันเพื่อทำการพยากรณ์ในชุดข้อมูล

จากนั้น Meta Learner ซึ่งเป็นแบบจำลองขั้นสุดท้ายจะทำการเรียนรู้จากผลลัพธ์ที่ได้จาก Base Learners โดยทำการปรับผลลัพธ์เหล่านี้เพื่อสร้างการพยากรณ์ที่ถูกต้องมากยิ่งขึ้น Meta Learner มีบทบาทในการผสมผสานผลลัพธ์จากแบบจำลองเบื้องต้นเหล่านี้ให้มีความแม่นยำมากขึ้นผ่านการเรียนรู้จากความผิดพลาดที่อาจเกิดขึ้นในแต่ละแบบจำลองเดี่ยว โดยเทคนิคนี้มีข้อได้เปรียบที่สำคัญเมื่อเทียบกับวิธีการ Voting, Bagging, หรือ Boosting เนื่องจาก Stacking สามารถรวมแบบจำลองที่ถูกสร้างจากอัลกอริทึมที่แตกต่างกันอย่างมีประสิทธิภาพ ช่วยให้สามารถรวมจุดแข็งของแบบจำลองแต่ละแบบและลดจุดอ่อนที่อาจเกิดขึ้นจากการใช้แบบจำลองเพียงหนึ่งเดียว ทำให้เกิดความเสถียรและความแม่นยำในการพยากรณ์ที่ดียิ่งขึ้น (Dietterich, 2000) ซึ่งรายละเอียดของการวัดประสิทธิภาพที่ใช้ มีดังนี้

1. Accuracy คือ การวัดความถูกต้องของแบบจำลอง โดยคำนวณจากสัดส่วนของการพยากรณ์ที่ถูกต้องทั้งหมด รวมทั้ง True Positive และ True Negative ต่อจำนวนข้อมูลทั้งหมด โดยค่าที่สูงใกล้ 100% แสดงถึงความแม่นยำของแบบจำลองในการพยากรณ์ผลลัพธ์ ดังสมการที่ 1

$$\text{Accuracy} = \frac{\text{TP+T}}{\text{TP+FP+TN+F}} \quad (1)$$

โดยที่

True Positive (TP) คือ การที่แบบจำลองทำนายว่าข้อมูลเป็นจริง และตรงกับความจริง

True Negative (TN) คือ การที่แบบจำลองทำนายว่าข้อมูลไม่เป็นจริง และตรงกับความจริง

False Positive (FP) คือ การที่แบบจำลองทำนายว่าข้อมูลเป็นจริงแต่ไม่ตรงกับความจริง

False Negative (FN) คือ การที่แบบจำลองทำนายว่าข้อมูลไม่เป็นจริงแต่ไม่ตรงกับความจริง

2. AUC คือ ค่าที่บ่งบอกถึงความสามารถของแบบจำลองในการแยกแยะระหว่างคลาสต่าง ๆ โดยมีค่าอยู่ระหว่าง 0 และ 1 ซึ่งค่าเข้าใกล้ 1 หมายถึงประสิทธิภาพดีมาก

3. F-measure คือ ค่าที่ใช้ประเมินประสิทธิภาพของแบบจำลองการจำแนกประเภท โดยรวมการพิจารณาทั้งความแม่นยำ (precision) และความสามารถในการตรวจจับ (recall) ซึ่งค่าที่สูงบ่งชี้ถึงความสมดุลที่ดีระหว่างการระบุผลลัพธ์ที่เป็นบวก และการลดความผิดพลาดจากการพลาดการตรวจพบ (false negatives) และผลบวกลวง (false positives) ดังสมการที่ 2

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

4. Root Mean Square Error (RMSE) คือ รากที่สองของค่าเฉลี่ยของค่าความคลาดเคลื่อนที่ยกกำลังสอง Squared Errors ทั้งหมด ดังสมการที่ 3

$$\text{RMSE} = \frac{1}{\text{n}}\sqrt{\sum_{\text{i=1}}^{\text{n}}(\text{y}_{\text{i}}-\widehat{\text{y}_{\text{i}}})^2} \quad (3)$$

5. Mean Squared Error (MSE) คือ ค่าความคลาดเคลื่อนระหว่างค่าที่พยากรณ์กับค่าจริง ที่ยกกำลังสอง ดังสมการที่ 4

$$\text{MSE} = \frac{1}{\text{n}}\sum_{\text{i=1}}^{\text{n}}(\text{y}_{\text{i}}-\widehat{\text{y}_{\text{i}}})^2 \quad (4)$$

โดยที่

n คือ จำนวนตัวอย่าง

y_i คือ ค่าจริง

$\hat{y}_i$ คือ ค่าที่พยากรณ์

ถ้าหากค่า RMSE และ MSE ต่ำหรือเข้าใกล้ 0 แสดงว่าแบบจำลองมีประสิทธิภาพสูงในการพยากรณ์

ส่วนค่าพารามิเตอร์ที่กำหนดให้แต่ละแบบจำลองการพยากรณ์นั้นใช้การปรับค่าพารามิเตอร์แบบอัตโนมัติโดยใช้เครื่องมือ Optimize Parameters

ประเภท Regression การปรับค่าพารามิเตอร์ของแต่ละอัลกอริทึม สำหรับ Artificial Neural Network ใช้ learning rate เป็นค่าที่กำหนดความเร็วในการปรับปรุงค่า ในการปรับค่า Max ค่า Min ให้อยู่ในช่วงที่ 0 - 1 ปรับค่าทีละ 1 training cycles คือ จำนวนรอบ ในการปรับค่า Max ค่า Min ให้อยู่ในช่วงที่ 1 - 100 ปรับค่าทีละ 10 hidden layer size 5 จำนวน 2 layer ส่วน Support Vector Machine ปรับ Parameter เป็น Anova เป็นฟังก์ชันที่สามารถใช้ในการทำให้ข้อมูลสามารถแยกออกจากกันได้ในกรณีที่ข้อมูลไม่ได้เป็นเชิงเส้น ใช้ C ปรับค่าเพื่อลดข้อผิดพลาดจากข้อมูลฝึก ในการปรับค่า Max ค่า Min ให้อยู่ในช่วง 0.1 – 10 ใช้ max iterations ใช้กำหนดจำนวนครั้งสูงสุดที่อัลกอริทึมจะทำการปรับค่าพารามิเตอร์ปรับค่า Max ค่า Min ให้อยู่ในช่วง 1 - 100 ส่วน Generalized Linear Model ปรับ Parameter ในช่อง stopping rounds เป็นจำนวนรอบที่โมเดลจะหยุดการฝึกเป็น 3 รอบ ช่อง stopping tolerance เป็น 0.001 และใช้ standardize ปรับสเกลของฟีเจอร์ให้อยู่ในช่วงที่เท่ากัน ในช่อง Value List ใช้ True

ส่วนประเภท Classification การปรับค่าพารามิเตอร์ของแต่ละอัลกอริทึม สำหรับ Decision Tree ใช้ maximal depth ความลึกสูงสุดของต้นไม้ ปรับค่า Max ค่า Min ให้อยู่ในช่วง 1 – 10 Criterion ใช้ในการแยกข้อมูลใดของแต่ละโหนดในช่อง Value List ใช้ gain ratio, information gain, gini index และ accuracy ส่วน Gradient Booting Tree ใช้ maximum number of threads จำนวนเธรดสูงสุดที่สามารถใช้ในระหว่างการประมวลผล ในการปรับค่า Max ค่า Min ให้อยู่ช่วงที่ 1 – 100 ใน min rows จำนวนข้อมูลขั้นต่ำที่ต้องมีในแต่ละโหนดเพื่อทำการ split ข้อมูลในโหนดนั้นๆ การปรับค่า Max ค่า Min ให้อยู่ช่วงที่ 1 – 10 ส่วน Random Forest ใช้ maximal depth ความลึกสูงสุดที่ต้นไม้จะเติบโตได้ในแต่ละต้น ปรับค่า Max ค่า Min ให้อยู่ในช่วง 1 – 10 maximum number of threads เธรดสูงสุดที่สามารถใช้ในการประมวลผล ในการปรับค่า Max ค่า Min ให้อยู่ช่วงที่ 1 – 100 และ Criterion ใช้ในการแยกข้อมูลใดของแต่ละโหนด ในช่อง Value List ใช้ gain ratio, information gain, gini index และ accuracy

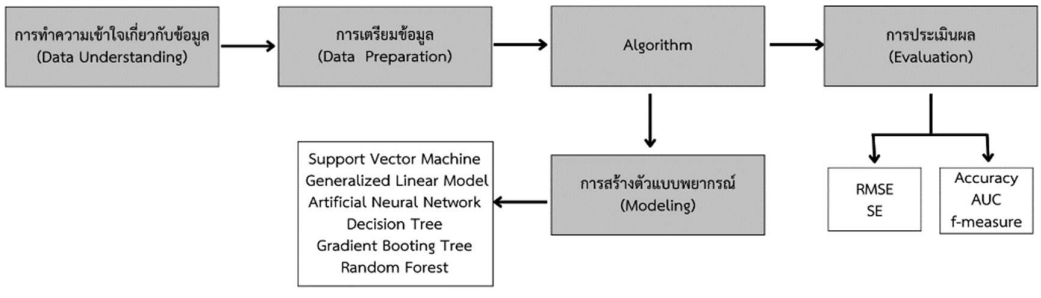


Figure 4 Forecasting model development process

Table 2 Datasets used in the study.

Data set	Details	Note
Monthly age-group data (age-month)	Month, Year, Annual population of Mueang Phayao District, Average monthly temperature, Average monthly rainfall, Average monthly relative humidity, Number of rainy days per month, Age group, Distance between outbreak areas and water sources, Distance between repeated outbreak areas, Forest area.	(In the regression analysis, the label used is the number of patients per 100,000 population per month or per week; in the classification analysis, the label used is the presence or absence of patients.)
Monthly gender-group data (sex-month)	Month, Year, Annual population of Mueang Phayao District, Average monthly temperature, Average monthly rainfall, Average monthly relative humidity, Number of rainy days per month, Gender group, Distance between outbreak areas and water sources, Distance between repeated outbreak areas, Forest area.	

ชุดข้อมูล	รายละเอียด	หมายเหตุ
Monthly sub-district data (subdistrict-month)	Subdistrict, Month, Year, Annual population by subdistrict, Average monthly temperature, Average monthly rainfall, Average monthly relative humidity, Number of rainy days per month, Population by subdistrict for each year, Distance between outbreak areas and water sources, Distance between repeated outbreak areas, Forest area.	(In the regression analysis, the label used is the number of patients per 100,000 population per month or per week; in the classification analysis, the label used is the presence or absence of patients.)
Weekly subdistrict data (subdistrict-week)	Subdistrict, Week, Year, Annual population by subdistrict, Average weekly temperature, Average weekly rainfall, Average weekly relative humidity, Number of rainy days per week, Distance between outbreak areas and water sources, Distance between repeated outbreak areas in the subdistrict health area, Forest area.	
Weekly age-group data (age-week)	Week, Year, Annual population of Mueang Phayao District, Average weekly temperature, Average weekly rainfall, Average weekly relative humidity, Number of rainy days per week, Age group, Distance between outbreak areas and water sources, Distance between repeated outbreak areas, Forest area.	
Weekly gender-disaggregated data (sex-week)	Week, Year, Annual population of Mueang Phayao District, Average weekly temperature, Average weekly rainfall, Average weekly relative humidity, Number of rainy days per week, Gender group, Distance between outbreak areas and water sources, Distance between repeated outbreak areas, Forest area.	

ผลการทดลอง

ผลการวัดประสิทธิภาพของแบบจำลองการพยากรณ์ประเภท Regression ซึ่งในชุดข้อมูลมีปัจจัยด้านระบบสารสนเทศภูมิศาสตร์ โดยใช้ Support Vector Machine, Generalized Linear Model และ Artificial Neural Network พบว่าแบบจำลอง Artificial Neural Network มีประสิทธิภาพสูงสุด ในชุดข้อมูล sex-week ซึ่งให้ค่า RMSE ต่ำสุดที่ 2.350 ดังแสดงใน Table 3

ผลการวัดประสิทธิภาพของแบบจำลองการพยากรณ์ประเภท Regression ซึ่งในชุดข้อมูลไม่ได้นำปัจจัยด้านระบบสารสนเทศภูมิศาสตร์มาใช้ในชุดข้อมูล โดยใช้ Support Vector Machine, Generalized Linear Model และ Artificial Neural Network พบว่าแบบจำลอง Artificial Neural Network มีประสิทธิภาพสูงสุด ในชุดข้อมูล sex-week ซึ่งให้ค่า RMSE ต่ำสุดที่ 0.968 ดังแสดงใน Table 4

Table 3 Performance measurement results of Regression with geographic information system factors

Datasets	Support Vector Machine		Generalized Linear Model		Artificial Neural Network	
	RMSE	MSE	RMSE	MSE	RMSE	MSE
age-month	8.708	79.177	23.904	620.361	14.374	208.107
sex-month	8.008	70.643	14.093	246.917	9.248	100.780
subdistrict-month	41.001	1681.967	100.817	11317.320	34.311	1347.955
subdistrict-week	22.158	524.983	43.522	1918.165	13.904	193.514
age-week	3.346	11.521	9.557	92.275	5.720	32.767
sex-week	2.389	5.716	5.163	27.073	2.350	5.542

Table 4 Performance measurement results of Regression

Datasets	Support Vector Machine		Generalized Linear Model		Artificial Neural Network	
	RMSE	MSE	RMSE	MSE	RMSE	MSE
age-month	5.921	37.412	5.211	27.593	3.925	15.687
sex-month	3.097	11.146	2.942	9.146	2.149	4.772
subdistrict-month	77.368	6564.470	71.536	5216.128	62.428	4303.912
subdistrict-week	31.291	982.351	4.913	26.385	5.079	27.350
age-week	8.127	66.162	1.889	3.600	1.876	3.596
sex-week	6.166	38.058	1.027	1.064	0.968	0.942

ผลการวัดประสิทธิภาพของแบบจำลองการพยากรณ์ประเภท Classification ซึ่งในชุดข้อมูลมีปัจจัยด้านระบบสารสนเทศภูมิศาสตร์ โดยใช้ Decision Tree, Gradient Boosting Tree และ Random Forest พบว่า Random Forest มีความแม่นยำสูงสุดในทุกชุดข้อมูล โดยเฉพาะในชุดข้อมูล subdistrict-month ซึ่งให้ค่า Accuracy ที่ 99.82% ดังแสดงใน Table 5

ผลการวัดประสิทธิภาพของแบบจำลองการพยากรณ์ประเภท Classification ซึ่งในชุดข้อมูลไม่ได้นำปัจจัยด้านระบบสารสนเทศภูมิศาสตร์มาใช้ในชุดข้อมูล โดยใช้ Decision Tree, Gradient Boosting Tree และ Random Forest พบว่า Gradient Boosting Tree มีความแม่นยำสูงสุดในทุกชุดข้อมูล โดยเฉพาะในชุดข้อมูล subdistrict-week ซึ่งให้ค่า Accuracy ที่ 99.62% ดังแสดงใน Table 6

Table 5 Performance measurement results of Classification with geographic information system factors

Datasets	Decision Tree			Gradient Booting Tree			Random Forest		
	Accuracy	AUC	F-measure	Accuracy	AUC	F-measure	Accuracy	AUC	F-measure
age-month	65.99%	0.695	53.44%	85.29%	0.904	84.84%	88.60%	0.923	88.22%
sex-month	62.50%	0.576	45.37%	81.67%	0.853	79.41%	75.00%	0.755	72.53%
subdistrict-month	98.13%	0.989	98.18%	98.89%	0.999	98.89%	99.82%	1.000	99.83%
subdistrict-week	63.79%	0.733	56.20%	90.37%	0.955	90.95%	95.48%	0.996	95.70%
age-week	70.12%	0.727	75.68%	89.37%	0.946	89.29%	90.63%	0.983	91.08%
sex-week	78.33%	0.849	74.76%	88.67%	0.954	89.36%	91.13%	0.962	91.34%

Table 6 Performance measurement results of Classification

Datasets	Decision Tree			Gradient Booting Tree			Random Forest		
	Accuracy	AUC	F-measure	Accuracy	AUC	F-measure	Accuracy	AUC	F-measure
age-month	89.01%	0.855	92.36%	85.71%	0.860	90.05%	85.25%	0.778	89.76%
sex-month	86.90%	0.964	82.87%	85.00%	0.907	80.00%	78.33%	0.708	73.95%
subdistrict-month	97.22%	0.963	97.30%	98.98%	0.992	98.98%	92.92%	0.949	93.34%
subdistrict-week	97.62%	0.757	97.75%	99.62%	0.998	99.48%	96.84%	0.982	96.95%
age-week	90.33%	0.942	91.10%	88.80%	0.962	88.62%	84.39%	0.877	85.65%
sex-week	85.11%	0.917	81.86%	85.96%	0.923	86.36%	77.14%	0.801	78.98%

ในส่วนของ Stacking model ผู้วิจัยได้สลับทั้ง 3 อัลกอริทึมในการสร้างแบบจำลอง โดยใช้ 2 อัลกอริทึมเป็น Base Learner และใช้อีก 1 อัลกอริทึมเป็น Stacking Model Learner จากการสร้างแบบจำลองประเภท Regression ผลการศึกษาพบว่าการใช้ Generalized Linear Model เป็น Stacking Model Learner ให้ประสิทธิภาพดีที่สุด ในชุดข้อมูล age-week มีค่า RMSE ที่ 1.457 ดังแสดงใน Table 7 ส่วนผลของการสลับ Stacking Model Learner แสดงดัง Table 8 และ 9

Table 7 Regression performance measurement results with Generalized Linear Model in Stacking Model Learner

Datasets	Regression - Generalized Linear Model	
	RMSE	MSE
age-month	12.513	159.305
sex-month	8.970	85.451
subdistrict-month	18.518	344.392
subdistrict-week	11.739	138.044
age-week	1.457	2.122
sex-week	3.438	11.836

Table 8 Regression performance measurement results with Support Vector Machine in Stacking Model Learner

Datasets	Regression - Support Vector Machine	
	RMSE	MSE
age-month	8.686	76.257
sex-month	8.136	69.249
subdistrict-month	26.373	761.848
subdistrict-week	27.927	800.542
age-week	4.553	20.872
sex-week	2.162	4.700

Table 9 Regression performance measurement results with Artificial Neural Network in Stacking Model Learner

Datasets	Regression - Artificial Neural Network	
	RMSE	MSE
age-month	8.835	84.318
sex-month	8.711	80.090
subdistrict-month	22.538	517.055
subdistrict-week	19.369	399.752
age-week	6.310	41.419
sex-week	2.231	5.085

จากการสร้างแบบจำลองประเภท Classification ในส่วนของ Stacking model ซึ่งผู้วิจัยได้สลับทั้ง 3 อัลกอริทึมในการสร้างแบบจำลอง ซึ่งผลการศึกษาพบว่าการใช้ Gradient Booting Tree เป็น Stacking Model Learner ให้ประสิทธิภาพดีที่สุด ในชุดข้อมูล subdistrict-month มีค่า Accuracy ที่ 99.12% ดังแสดงใน Table 10 ส่วนผลของการสลับ Stacking Model Learner แสดงดัง Table 11 และ 12

นอกจากนี้ ผู้วิจัยได้ทำการประยุกต์ใช้แบบจำลองการในการพยากรณ์ โดยได้ทำการเลือกแบบจำลองที่มีประสิทธิภาพสูงสุดในส่วนของ Stacking Model ที่ใช้ Generalized Linear Model เป็น Stacking Model Learner มาใช้ในการพยากรณ์จำนวนผู้ป่วยโรคไข้เลือดออกในอำเภอเมืองพะเยา จังหวัดพะเยา โดยใช้ข้อมูลปี 2566 โดยผลลัพธ์จากพยากรณ์นี้ได้ถูกนำเสนอในรูปแบบแผนที่ ดังแสดงใน Figure 5 เพื่อแสดงข้อมูลเชิงพื้นที่อย่างชัดเจน ซึ่งช่วยให้ผู้มีส่วนเกี่ยวข้องสามารถนำข้อมูลไปใช้ในการวางแผน และตัดสินใจได้อย่างมีประสิทธิภาพมากยิ่งขึ้น

Table 10 Classification performance measurement results with Gradient Booting Tree in Stacking Model Learner

Datasets	Classification - Gradient Booting Tree					
	Accuracy	AUC	Precision	Recall	F-measure	Specificity
age-month	88.33%	0.893	83.92%	93.67%	88.40%	83.43%
sex-month	98.33%	1.000	100.00%	95.56%	97.50%	100.00%
subdistrict-month	99.12%	0.991	98.55%	99.78%	99.15	98.44%
subdistrict-week	96.27%	0.963	93.09%	100.00%	96.42%	92.50%
age-week	95.06%	0.953	91.05%	99.38%	95.03%	91.03%
sex-week	91.13%	0.910	87.66%	96.70%	91.94%	85.02%

Table 11 Classification performance measurement results with Random Forest in Stacking Model Learner

Datasets	Classification - Random Forest					
	Accuracy	AUC	Precision	Recall	F-measure	Specificity
age-month	88.85%	0.925	88.60%	88.35%	88.32%	89.33%
sex-month	72.50%	0.749	77.39%	61.97%	68.58%	82.05%
subdistrict-month	98.95%	1.000	97.99%	100.00%	98.98%	97.84%
subdistrict-week	96.59%	0.996	93.68%	100.00%	96.73%	93.15%
age-week	91.52%	0.980	84.98%	100.00%	91.86%	83.75%
sex-week	90.89%	0.958	92.67%	89.60%	91.09%	92.29%

Table 12 Classification performance measurement results with Decision Tree in Stacking Model Learner

Datasets	Classification - Decision Tree					
	Accuracy	AUC	Precision	Recall	F-measure	Specificity
age-month	82.00%	0.840	79.82%	85.12%	81.89%	79.18%
sex-month	75.83%	0.560	78.88%	72.58%	73.84%	78.97%
subdistrict-month	98.89%	0.988	97.99%	99.89%	98.93%	97.84%
subdistrict-week	86.43%	0.899	80.06%	97.47%	87.85%	75.29%
age-week	86.43%	0.883	80.99%	93.85%	86.88%	79.64%
sex-week	88.18%	0.906	87.30%	90.54%	88.83%	85.61%

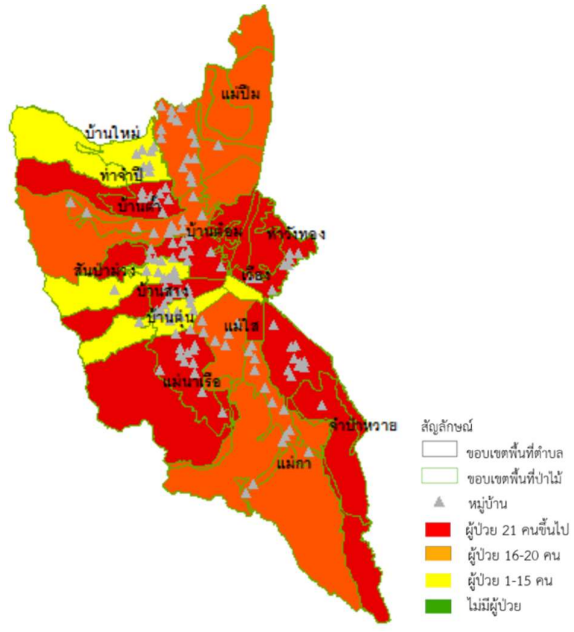


Figure 5 Map displaying the dengue fever outbreak forecast for 2023, by sub-district

อภิปรายผลการทดลอง

จากการศึกษาการใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) ในการพยากรณ์โรคไข้เลือดออกร่วมกับระบบสารสนเทศทางภูมิศาสตร์ (GIS) ในการพยากรณ์โรคไข้เลือดออกในอำเภอเมืองพะเยา จังหวัดพะเยา ผู้วิจัยได้ศึกษาวิเคราะห์ถึงความสัมพันธ์ของปัจจัยต่าง ๆ ในการเกิดโรคไข้เลือดออก โดยการใช้เทคนิคการเรียนรู้ของเครื่องสามารถนำไปใช้เป็นเครื่องมือในการเฝ้าระวัง สามารถควบคุมโรคได้อย่างมีประสิทธิภาพในอนาคต การพยากรณ์ที่แม่นยำ และมีประสิทธิภาพจะช่วยให้หน่วยงานที่เกี่ยวข้องสามารถวางแผนและดำเนินการป้องกันการระบาดของโรคได้ทันท่วงที ซึ่งจะช่วยลดจำนวนผู้ป่วย หรือผลกระทบทางสังคมที่เกิดจากการระบาดของโรคไข้เลือดออก ซึ่งการใช้ระบบสารสนเทศภูมิศาสตร์ช่วยให้สามารถวิเคราะห์ และจัดการข้อมูลภูมิศาสตร์ที่มีความซับซ้อนได้ เช่น ระยะห่างระหว่างพื้นที่การระบาดกับแหล่งน้ำและพื้นที่ป่าไม้ ซึ่งเป็นปัจจัยที่มีผลต่อการแพร่ระบาดของโรคไข้เลือดออก การนำข้อมูลเหล่านี้มาวิเคราะห์ร่วมกับข้อมูลทางอุตุนิยมวิทยา และข้อมูลประชากรช่วยให้สามารถสร้างแบบจำลองที่มีความแม่นยำในการพยากรณ์การระบาดของโรคได้ โดยผู้วิจัยได้นำปัจจัยดังกล่าวมาสร้างเป็นแบบรายเดือน และรายสัปดาห์

การวิเคราะห์ปัจจัยที่ส่งผลต่อการพยากรณ์ ผู้วิจัยได้สร้างชุดข้อมูลที่ใช้และไม่ใช้ระบบสารสนเทศทางภูมิศาสตร์ เพื่อเปรียบเทียบข้อแตกต่างของชุดข้อมูลพบว่า ชุดข้อมูลที่ใช้ระบบสารสนเทศภูมิศาสตร์ ในประเภท Regression พบว่า แบบจำลอง ที่มีประสิทธิภาพสูงสุด อยู่ในชุดข้อมูล sex-week โดยให้ค่า RMSE ต่ำสุดที่ 2.350 ส่วนชุดข้อมูลที่ไม่ใช้ ให้ค่า RMSE ต่ำสุดที่ 0.968 ส่วนในประเภท Classification ชุดข้อมูลที่ใช้ระบบสารสนเทศภูมิศาสตร์ พบว่าแบบจำลอง ที่มีประสิทธิภาพสูงสุด อยู่ในชุดข้อมูล subdistrict-month โดยให้ค่า Accuracy ที่ 99.82% ส่วนชุดข้อมูลที่ไม่ใช้ ให้ค่า Accuracy ที่ 99.62% แสดงถึงผลของปัจจัยด้านภูมิสารสนเทศมีผลต่อประสิทธิภาพของแบบจำลอง

จากการเปรียบเทียบอัลกอริทึมและชุดข้อมูล พบว่า ในประเภท Regression พบว่า Generalized Linear Model ให้ประสิทธิภาพดีที่สุด ที่ RMSE 1.457 ในชุดข้อมูล sex-week ส่วนผลที่ได้ประเภท Classification พบว่า Gradient Booting Tree ให้ประสิทธิภาพดีที่สุด ที่ Accuracy 99.12% ในชุดข้อมูล subdistrict-month เป็นแบบจำลองที่เหมาะสมที่สุดในการพยากรณ์การระบาดของโรคไข้เลือดออก เนื่องจากสามารถให้ค่าความแม่นยำ และประสิทธิภาพสูงสุดในการพยากรณ์ ซึ่งสอดคล้องกับงานวิจัยวิธีการที่ขับเคลื่อนด้วยข้อมูลในการพยากรณ์ และเฝ้าระวังโรคไข้เลือดออก ซึ่งแบบจำลองที่มีประสิทธิภาพสูงสุดสองแบบคือ Neural Networks และ Decision Trees ที่ Accuracy 52% (Sylvestre *et al*., 2022) งานวิจัยแบบจำลองการพยากรณ์ของการระบาดโรคไข้เลือดออก พบว่าวิธี Decision Tree เป็นแบบจำลองในการสร้างแบบจำลองสำหรับการพยากรณ์การระบาดของโรคไข้เลือดออกที่มีประสิทธิภาพสูงสุดโดยให้ค่าความแม่นยำสูงสุด 69.83% และ 75.4% ตามลำดับ (จิรโรจน์ ตอสะสุกุล, 2564) งานวิจัยการพยากรณ์จำนวนผู้ป่วยโรคไข้เลือดออกในประเทศมาเลเซีย พบว่า SSA-LSTM ซึ่งเป็นแบบจำลองที่ใช้ทั้ง stacked LSTM layers และ spatial attention ทำผลงานได้ดีที่สุด โดยมีค่า RMSE เท่ากับ 3.17 (Majeed *et al*., 2023) และงานวิจัยการพยากรณ์พื้นที่เสี่ยงต่อการเกิดโรคไข้เลือดออก เพื่อใช้ในการเฝ้าระวัง และวางแผนป้องกันใน โดยการใช้แบบสำรวจลูกน้ำยุงลาย เช่น House Index (HI), Container Index (CI) และ Breteau Index (BI) เพื่อนำมาพัฒนาแบบจำลองการวิเคราะห์และพยากรณ์พื้นที่เสี่ยงในระบบ GIS ซึ่งในงานวิจัยนี้ไม่ได้ใช้ Machine Learning ในการพยากรณ์ (สุรศักดิ์ สุขสาย และวนิดา แก่นอากาศ, 2550)

จากผลการศึกษาผลว่า Generalized Linear Model และ Gradient Boosting Tree มีประสิทธิภาพที่ดีที่สุดในการพยากรณ์ เนื่องจากแบบจำลองเหล่านี้มีความสามารถในการจับความสัมพันธ์ที่ซับซ้อนในข้อมูล สามารถจัดการกับความแปรปรวน ความเอนเอียงในข้อมูลได้อย่างมีประสิทธิภาพ ส่งผลให้การพยากรณ์มีความแม่นยำสูง และสามารถใช้เป็น

เครื่องมือที่มีประสิทธิภาพในการคาดการณ์การระบาดของโรคได้ สามารถจัดการกับความไม่สมดุลของข้อมูล และเพิ่มประสิทธิภาพการพยากรณ์ได้เป็นอย่างดี ทำให้สามารถคาดการณ์การระบาดของโรคได้อย่างแม่นยำ

จากการศึกษาปัจจัยที่ใช้ในการสร้างชุดข้อมูล ผู้แต่งได้ใช้ปัจจัยที่คาดว่าจะมีผลต่อการระบาด เช่นปัจจัยที่ส่งผลต่อยุงที่เป็นพาหะไข้เลือดออก ซึ่งอาจจะเป็นแหล่งน้ำที่เป็นที่วางไข่ ปริมาณน้ำฝน ความชื้นหรือป่าไม้ซึ่งอาจมีผลต่อการเจริญเติบโตของยุง และปัจจัยด้านอื่น ที่เกี่ยวข้อกับการระบาด เช่น จำนวนประชากร การเกิดการระบาดซ้ำในพื้นที่ และปัจจัยที่ใช้ในการจัดกลุ่มชุดข้อมูล เช่น เพศ อายุ ฯลฯ ทั้งนี้การเกิดการระบาดเป็นกระบวนการธรรมชาติที่ซับซ้อน จำเป็นต้องมีการศึกษาเพิ่มเติมถึงปัจจัยที่ใช้ต่อไป อีกทั้งงานวิจัยนี้ยังมีข้อจำกัดบางประการ เช่น การรวบรวมข้อมูลจากแหล่งข้อมูลที่หลากหลาย ซึ่งอาจทำให้เกิดความไม่สมบูรณ์ และความไม่สม่ำเสมอของข้อมูลในชุดข้อมูลรวมถึงความต้องการของแบบจำลองการเรียนรู้ของเครื่องที่ต้องอาศัยข้อมูลจำนวนมาก และการปรับแต่งแบบจำลองให้มีความเหมาะสม ซึ่งอาจเป็นข้อจำกัดในการนำไปใช้ในบริบทที่มีข้อจำกัดด้านข้อมูล และอาจเกิดการ Overfitting ได้ ดังนั้น จำเป็นต้องมีการวิจัยเพิ่มเติมเพื่อพัฒนากับปรับปรุงแบบจำลองให้มีความยืดหยุ่น และมีประสิทธิภาพมากยิ่งขึ้นในสภาพแวดล้อมที่หลากหลาย

กิตติกรรมประกาศ

ผู้วิจัยขอขอบคุณโรงพยาบาลพะเยา สถานีอุตุนิยมวิทยาพะเยา และสำนักงานสาธารณสุขจังหวัดพะเยา ที่ให้ความอนุเคราะห์ข้อมูล ให้คำปรึกษา และข้อเสนอแนะในการทำวิจัยครั้งนี้

เอกสารอ้างอิง

กรมควบคุมโรค. (2560). *ความรู้ทั่วไปโรคไข้เลือดออก*. กระทรวงสาธารณสุข.

กรมควบคุมโรค. (2563). *ไข้เลือดออก (Dengue Fever และ Dengue Hemorrhagic Fever) (ICD 10 A90, A91)*. https://ddc.moph.go.th/dcd/pagecontent.php?page=565&dept=dcd

กอบเกียรติ สระอุบล. (2563). *เรียนรู้ Data Science และ AI: Machine Learning ด้วย Python* (พิมพ์ครั้งที่ 1). มีเดียเนทเวิร์ค.

จิรโรจน์ ตอสะสุกุล. (2564). แบบจำลองการพยากรณ์ของการระบาดโรคไข้เลือดออกโดยใช้เทคนิคการทำเหมืองข้อมูล. *วารสารวิชาการซายน์เทค มรภ.ภูเก็ต*, *5*(2), 51–60.

โรงพยาบาลพญาไท. (ม.ป.ป.). *ไข้เลือดออก อย่าปล่อยให้เป็นซ้ำ อันตรายถึงชีวิต*. https://www.phyathai-sriracha.com/article/medicine/

สายชล สินสมบูรณ์ทอง. (2560). *การทำเหมืองข้อมูล เล่ม 1: การค้นหาความรู้จากข้อมูล*. จามจุรีโปรดักส์.

สำนักงานสาธารณสุขจังหวัดพะเยา. (2566). *บทสรุปผู้บริหาร สถานการณ์โรคไข้เลือดออก จังหวัดพะเยา (ฉบับที่ 1 ปี 2566)*.

สุรศักดิ์ สุขสาย และ วนิดา แก่นอากาศ. (2550). การพยากรณ์พื้นที่เสี่ยงต่อการเกิดโรคไข้เลือดออกเพื่อการวางแผนเฝ้าระวังและป้องกันโรคในจังหวัดอุบลราชธานี. *วารสารวิจัย มข. (ฉบับบัณฑิตศึกษา)*, *7*(2), 83–97. https://ph02.tci-thaijo.org/index.php/gskku/article/view/23776/20224

ศูนย์วิจัยภูมิสารสนเทศเพื่อประเทศไทย. (ม.ป.ป.). *ความหมายของคำว่า "ระบบสารสนเทศภูมิศาสตร์"*. https://www.gisthai.org/about-gis/gis.html

Dietterich, T. G. (2000). Ensemble methods in machine learning. In J. Kittler & F. Roli (Eds.), *Multiple classifier systems* (pp. 1–15). Springer. https://doi.org/10.1007/3-540-45014-9_1

EFD Initiative. (2024, September 5). *Forest cover could reduce hospital admissions caused by dengue fever*. https://www.efdinitiative.org/news/forest-cover-could-reduce-hospital-admissions-caused-dengue-fever

Hofmann, M., & Klinkenberg, R. (Eds.). (2013). *RapidMiner: Data mining use cases and business analytics applications*. CRC Press.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, 1137–1145. https://doi.org/10.5555/1643031.1643047

Majeed, M. A., Shafri, H. Z. M., Zulkafli, Z., & Wayayok, A. (2023). A deep learning approach for dengue fever prediction in Malaysia using LSTM with spatial attention. *International Journal of Environmental Research and Public Health*, *20*(5), 4087. https://doi.org/10.3390/ijerph20054087

Sylvestre, E., Joachim, C., Cécilia-Joseph, E., Bouzillé, G., Campillo-Gimenez, B., Cuggia, M., & Cabié, A. (2022). Data-driven methods for dengue prediction and surveillance using real-world and Big Data: A systematic review. *PLoS Neglected Tropical Diseases*, *16*(1), e0010056. https://doi.org/10.1371/journal.pntd.0010056