แนวทางการเรียนรู้ของเครื่องเพื่อการทำนายไข้เลือดออก: กรณีศึกษาจังหวัดพะเยา

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

A Machine Learning Approach for Dengue Fever Prediction: Case Study of Phayao Province

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

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

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

Received: May 28, 2024; Revised: August 17, 2024; Accepted: August 20, 2024

บทคัดย่อ

โรคไข้เลือดออกเป็นโรคร้ายแรงที่เกิดจากเชื้อไวรัสโดยมียุงลายเป็นพาหะ เป็นปัญหาสำคัญของกระทรวงสาธารณสุขในหลายประเทศทั่วโลก งานวิจัยนี้จึงมีวัตถุประสงค์เพื่อศึกษาปัจจัยที่มีผลต่อการระบาดของโรคไข้เลือดออกและสร้างแบบจำลองในการพยากรณ์โรคไข้เลือดออกที่มีประสิทธิภาพ ด้วยเทคนิคการเรียนรู้ของเครื่อง โดยใช้ข้อมูลจากสถานีอุตุนิยมวิทยาพะเยา ได้แก่ สภาพภูมิอากาศ อุณหภูมิ ความชื้นสัมพัทธ์ ปริมาณน้ำฝน จำนวนวันที่ฝนตก และข้อมูลจำนวนประชากรในพื้นที่จังหวัดพะเยา จากฐานข้อมูล สำนักงานสาธารณสุขจังหวัดพะเยา รวมถึงจำนวนผู้ป่วยโรคไข้เลือดออก กลุ่มเพศ และกลุ่มอายุ จากโรงพยาบาลพะเยา ตั้งแต่ ปี พ.ศ. 2560 – 2565 โดยนำมาวิเคราะห์และสร้างแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่อง 10 วิธี โดยประเภท Regression ได้แก่ Support Vector Machines และ Linear Regression ส่วนประเภท Classification ได้แก่ Artificial Neural Network, Decision Tree, Naïve Bayes, K–Nearest Neighbors, Deep learning, Random Trees, Gradient Boosting และ Logistic Regression และทำการวัดประสิทธิภาพแบบจำลองด้วยวิธีการ 5–Fold Cross Validation

ข้อมูลทั้งหมดถูกสร้างเป็น dataset แบบรายเดือนและรายสัปดาห์ ซึ่งจากผลการวัดประสิทธิภาพของแบบจำลองประเภท Regression พบว่า Linear Regression ให้ประสิทธิภาพดีที่สุด ที่ RMSE 1.190 และในประเภท Classification พบว่า Deep learning เป็นแบบจำลองที่มีประสิทธิภาพที่สุด ที่ให้ประสิทธิภาพดีที่สุด ที่ Accuracy 99.84% โดยพิจารณาจากค่าความแม่นยำและค่าประสิทธิภาพโดยรวมสูงที่สุด ผลจากการทำวิจัยในครั้งนี้สามารถเป็นแนวทางในการประยุกต์ นำไปใช้ประโยชน์ได้อย่างแพร่หลายโดยเฉพาะหน่วยงานต่าง ๆ ที่เกี่ยวข้องในการวางแผนการเฝ้าระวังการหาพื้นที่เสี่ยงการระบาด การป้องกันและการควบคุมโรคไข้เลือดออกได้อย่างมีประสิทธิภาพ

คำสำคัญ: ไข้เลือดออก, เทคนิคการเรียนรู้ของเครื่อง, แบบจำลองการพยากรณ์, การถดถอยเชิงเส้น, การเรียนรู้เชิงลึก

[1] คณะเทคโนโลยีสารสนเทศและการสื่อสาร มหาวิทยาลัยพะเยา จังหวัดพะเยา 56000

Abstract

Dengue fever is a serious disease caused by a virus carried by Aedes mosquitoes. It is an important problem of ministries of health in many countries around the world. This research therefore aims to study factors affecting the outbreak of dengue fever and create an effective dengue fever prediction model using machine learning techniques. Data from the Phayao Meteorological Station, including climate, temperature, relative humidity, rainfall, number of rainy days and population data in Phayao Province was collected from the provincial public health database. In addition, the number of dengue fever patients, gender and age group collected from Phayao Hospital between 2017 and 2022 was analyzed and build a model with 10 machine learning techniques. Regression types include Support Vector Machines and Linear Regression. Classification types include Artificial Neural Network, Decision Tree, Naïve Bayes, K–Nearest Neighbors, Deep learning, Random Trees, Gradient Boosting, and Logistic Regression and measure model performance using the 5–Fold Cross Validation method.

All data is created in monthly and weekly datasets. Considering the highest overall accuracy and efficiency. From the performance measurement results of Regression, it was found that Linear Regression provides the best performance with an RMSE of 1.190. From the Classification results, Deep Learning was found to be the most effective model with the highest overall performance, reaching an Accuracy of 99.84%. The results from this research can be used as guidelines for application, especially by various agencies involved in surveillance planning to find areas at risk of spreading and to effectively prevent and control dengue fever.

Keywords: Dengue fever, Machine learning, Prediction model, Linear Regression, Deep Learning

บทนำ

ในปัจจุบันโรคไข้เลือดออกเป็นโรคร้ายแรงที่ติดต่อได้จากเชื้อไวรัสเดงกี ที่ถูกถ่ายทอดผ่านยุงลายบ้าน เป็นปัญหาสำคัญของกระทรวงสาธารณสุขในหลายประเทศทั่วโลก เนื่องจากมีการระบาดอย่างแพร่หลายและมีจำนวนผู้ป่วยเพิ่มขึ้น อย่างรวดเร็ว มากกว่า 300 ประเทศที่โรคไข้เลือดออกได้กลายเป็นโรคประจำถิ่น โดยมีผู้ป่วยสูงถึงร้อยละ 40 ของประชากรโลก โรคนี้มักพบในเขตร้อนและระบาดในช่วงฤดูฝนของทุก ๆ ปี โดยในแถบเอเชียมีผู้ป่วยร้อยละ 70 ของผู้ป่วยทั้งหมด ในประเทศไทยเริ่มมีการรายงานพบผู้ป่วยครั้งแรกในปี พ.ศ. 2492 และการระบาดใหญ่ที่สุดเกิดขึ้นในปี พ.ศ. 2530 โดยมีผู้ป่วยเพิ่มขึ้น อย่างมากเกิน 170,000 ราย และเกิดการเสียชีวิตเกิน 1,000 ราย ต่อมาประเทศไทยเริ่มพบผู้ป่วยเพิ่มขึ้นเป็นประจำ โดยในปีที่เกิดการระบาดใหญ่จะพบผู้ป่วยเกิน 100,000 รายและมีผู้เสียชีวิตเกิน 100 ราย การระบาดของโรคนี้ส่งผลกระทบต่อสุขภาพและเศรษฐกิจของประชากรอย่างมาก สรุปได้ว่า โรคไข้เลือดออกเป็นปัญหาร้ายแรงที่ต้องสนับสนุนและจัดการอย่างเร่งด่วน ในหลายประเทศทั่วโลก และการควบคุมการระบาดของโรคนี้เป็นเรื่องสำคัญที่ต้องการความร่วมมือของทุกภาคส่วนในสังคม (กรมควบคุมโรค, 2560)

โรคไข้เลือดออก (Dengue fever) เกิดจากเชื้อไวรัส โดยถูกส่งผ่านจากยุงลายที่เป็นพาหะ โดยการกัดของยุงลายที่เคยสัมผัสกับคนที่ติดเชื้อไวรัสเดงกี ไวรัสเดงกีมีสี่สายพันธุ์ (DENV–1 – DENV–4) ที่สามารถทำให้เกิดโรคไข้เลือดออกได้ ผู้ที่ติดเชื้อไวรัสอาจไม่มีอาการหรือมีอาการเล็กน้อย เช่น ไข้ขึ้นสูง ปวดศีรษะ ปวดกล้ามเนื้อ คลื่นไส้ อาเจียน และมีจ้ำเลือดหรือจุดเลือดออกบางจุดบนผิวหนัง ในกรณีที่โรคไข้เลือดออกเป็นรุนแรง อาจทำให้เกิดอาการรุนแรงเช่น ระบบ

ไหลเวียนเลือดล้มเหลว ช็อก หรือเสียชีวิตได้ ดังนั้น หากเสี่ยงต่อการติดเชื้อไวรัสโรคไข้เลือดออก ควรพบแพทย์เพื่อรับการตรวจและการรักษาโดยเร็ว (กรมควบคุมโรค, 2563)

การแพร่กระจายของไวรัสเดงกีเกิดจากยุงลายบ้าน (Aedes aegypti) และยุงลายสวน (Aedes albopictus) ซึ่งเป็นปัจจัยสำคัญที่ทำให้เกิดโรคไข้เลือดออก เนื่องจากยุงตัวเมียเป็นพาหะนำโรค โรคไข้เลือดออกสามารถแพร่กระจายได้อย่างรวดเร็ว การควบคุมและป้องกันโรคมีความสำคัญ เพราะมีปัจจัยต่าง ๆ ที่ ส่งผลให้เกิดโรคไข้เลือดออก เช่น จำนวนประชากรที่เพิ่มขึ้น สภาพอากาศที่เปลี่ยนแปลง อุณหภูมิ ปริมาณค่าน้ำฝน และความรู้ความเข้าใจของประชาชนในการป้องกันโรค ดังนั้น จึงต้องนำปัจจัยที่เกี่ยวข้องมาพยากรณ์โรคไข้เลือดออกเพื่อให้มีแนวทางการป้องกันล่วงหน้าที่มีประสิทธิภาพและลดอัตรา การเกิดโรคไข้เลือดออกในจังหวัดพะเยา

กระบวนการทำเหมืองข้อมูล (Data Mining) ประกอบด้วยขั้นตอนทั้งหมด 5 ขั้นตอน คือ การทำความเข้าใจข้อมูล โดยการวิเคราะห์ข้อมูลที่ได้จากการสำรวจเพื่อสร้างความคุ้นเคยกับข้อมูลและค้นหาความเข้าใจเชิงลึกเบื้องต้น การเตรียมข้อมูล เป็นการแปลงข้อมูลดิบให้อยู่ในรูปแบบที่เหมาะสมสำหรับการวิเคราะห์ การสร้างแบบจำลองการพยากรณ์ โดยการเลือกเทคนิคการสร้างตัวแบบให้เหมาะสม การประเมินผล ซึ่งเป็นการประเมินผลลัพธ์ตัวแบบที่ได้จากขั้นตอนการสร้างตัวแบบพยากรณ์สำหรับการวัดประสิทธิภาพที่สร้างขึ้นโดยใช้ข้อมูลมาทดสอบ และการนำไปใช้ โดยการนำตัวแบบที่ผ่านการประเมินและปรับแต่งแล้วมาลองปฏิบัติจริง เพื่อให้สามารถนำไปใช้ประโยชน์ได้จริง (สายชล สินสมบูรณ์ทอง, 2560)

การเรียนรู้ของเครื่อง (Machine Learning) เป็นการนำ Machine Learning มาประยุกต์ใช้ในการสร้างแบบจำลองทางคณิตศาสตร์ที่ช่วยให้คอมพิวเตอร์สามารถเรียนรู้จากชุดข้อมูลและประสบการณ์ในอดีต เพื่อทำการคาดการณ์และตัดสินใจโดยอัตโนมัติ โดยส่วนของโปรแกรมที่ทำการคำนวณประมวลผล การตั้งเงื่อนไข การตัดสินใจต่าง ๆ เรียกว่า Model ซึ่ง Model นี้จะทำหน้าที่เหมือนมันสมองของโปรแกรมที่สามารถรับ Input ประมวลผล และให้ Output ออกมาได้อย่างแม่นยำ โดยไม่จำเป็นต้องมีการป้อนคำสั่งจากผู้พัฒนาโดยตรง ทั้งนี้ ประสิทธิภาพของการทำนายหรือการตัดสินใจจะขึ้นอยู่กับปริมาณข้อมูลที่ใช้ในการฝึกอัลกอริทึม ยิ่งมีข้อมูลมากเท่าใด ประสิทธิภาพก็จะยิ่งสูงขึ้น (กอบเกียรติ สระอุบล, 2563)

จากรายงานการเฝ้าระวังโรค จากสำนักงานสาธารณสุขจังหวัดพะเยา วันที่ 1 มกราคม 2556 – 3 สิงหาคม 2566 มีรายงานผู้ป่วย 549 ราย อัตราป่วยเท่ากับ 115.62 ต่อประชากรแสนคน และมีผู้เสียชีวิต 1 ราย เปรียบเทียบในช่วงเวลาเดียวกัน พบว่าผู้ป่วยสูงกว่าปี 2565 เป็น 5.49 เท่า สูงกว่าค่ามัธยฐาน 5 ปีหลัง (ปี 2561 – 2565) เป็น 3.34 เท่า อัตราป่วยเป็นอันดับ ที่ 6 ของพื้นที่ 8 จังหวัดภาคเหนือ และอันดับที่ 26 ของประเทศ (สำนักงานสาธารณสุขจังหวัดพะเยา, 2566) ดังนั้นผู้วิจัยจึงทำการศึกษาโรคไข้เลือดออกโดยใช้เทคนิคการเรียนรู้ของเครื่องเพื่อสร้างแบบจำลองการพยากรณ์ ให้สามารถทำนายจำนวนผู้ป่วยโรคไข้เลือดออกในจังหวัดพะเยา ซึ่งจะสามารถหาแนวทางป้องกันล่วงหน้าได้อย่างมีประสิทธิภาพและช่วยลดโอกาสในการเกิดการระบาด

วัตถุประสงค์

1. เพื่อศึกษาปัจจัยที่มีผลต่อการระบาดของโรคไข้เลือดออก
2. เพื่อสร้างแบบจำลองในการพยากรณ์โรคไข้เลือดออก จังหวัดพะเยา
3. เพื่อทำนายจำนวนผู้ป่วยโรคไข้เลือดออกซึ่งจะนำไปใช้ในการวางแผนการเฝ้าระวัง ป้องกัน และควบคุมโรคไข้เลือดออกในอนาคตได้อย่างมีประสิทธิภาพ

งานวิจัยที่เกี่ยวข้อง

การศึกษาวิจัยเกี่ยวกับการประยุกต์ใช้เทคนิคการทำเหมืองข้อมูล (Data Mining) ในการหาแบบจำลองสำหรับการพยากรณ์การระบาดของโรคไข้เลือดออกที่มีความแม่นยำและประสิทธิภาพดีที่สุด ซึ่งผลการวิจัยที่ได้พบว่า Decision Tree เป็นเทคนิคที่เหมาะสมที่สุดจากการประเมินด้วยค่าความแม่น (Accuracy) และ ค่าประสิทธิภาพโดยรวม (F–measure) เท่ากับ 69.83% และ 75.4% ตามลำดับ (จิรโรจน์ ตอสะสุกุล, 2564)

การศึกษาโดยมีโมเดล Long Short–Term Memory (LSTM) หกแบบที่แตกต่างกันและเปรียบเทียบเพื่อการพยากรณ์ไข้เลือดออก ผลการทดลองพบว่า โมเดล SSA–LSTM ซึ่งใช้ทั้งเลเยอร์ LSTM ซ้อนกันและ spatial attention มีประสิทธิภาพดีที่สุด โดยมีค่าเฉลี่ยของ Root Mean Squared Error (RMSE) เท่ากับ 3.17 ในทุกช่วงเวลาการทำนายและยังทำนายได้ดีในรัฐอื่น ๆ โดยมีค่า RMSE 2.91 ถึง 4.55 (Majeed et al, 2023)

การศึกษาโดยใช้โมเดลการเรียนรู้ของเครื่อง LSTM (Long Short–Term Memory) ผลการวิจัยแสดงให้เห็นว่าโมเดล LSTM มีความแม่นยำในการพยากรณ์สูงสุด โดยค่า RMSE เท่ากับ 0.04 และ ค่า R–squared (R2) เท่ากับ 0.84 และวิธีการของการศึกษานี้คือ การนำโมเดล LSTM มาใช้ในการพยากรณ์การระบาดของไข้เลือดออกในรัฐคุชราตเป็นครั้งแรก โดยมีความแม่นยำสูงถึง 84% และสามารถนำไปใช้ในการวางแผนการจัดการสาธารณสุขในพื้นที่ได้อย่างมีประสิทธิภาพ (Mehta & Patel, 2023)

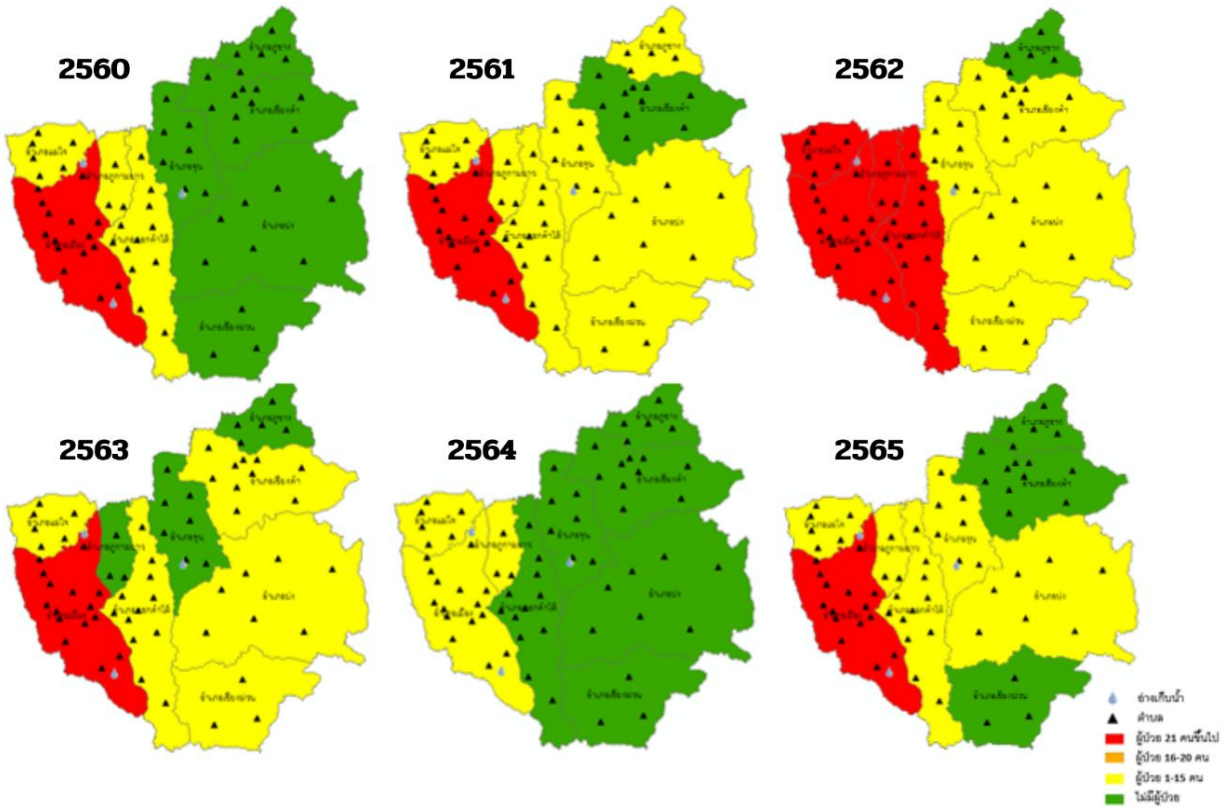
การศึกษาโดยใช้โมเดลการเรียนรู้ของเครื่องที่รวมข้อมูลทางอุตุนิยมวิทยาเพื่อตรวจสอบความสัมพันธ์ที่ซับซ้อนระหว่างการเปลี่ยนแปลงสภาพอากาศกับการแพร่ระบาดของโรค โดยมีการใช้อัลกอริธึมการเรียนรู้ของเครื่อง คือ XGBoost, Random Forest และ Support Vector Machine โดยที่โมเดล XGBoost ให้ผลการพยากรณ์ที่ดีที่สุด ค่า MAE = 89.12, RMSE = 156.07 และ R2 = 0.83 การศึกษานี้พบว่าปัจจัยทางอุตุนิยมวิทยา เช่น เวลา เมฆคลุม และปริมาณฝน มีบทบาทสำคัญในการแพร่ระบาดของไข้เลือดออก (Tian et al., 2024)

วิธีการศึกษา

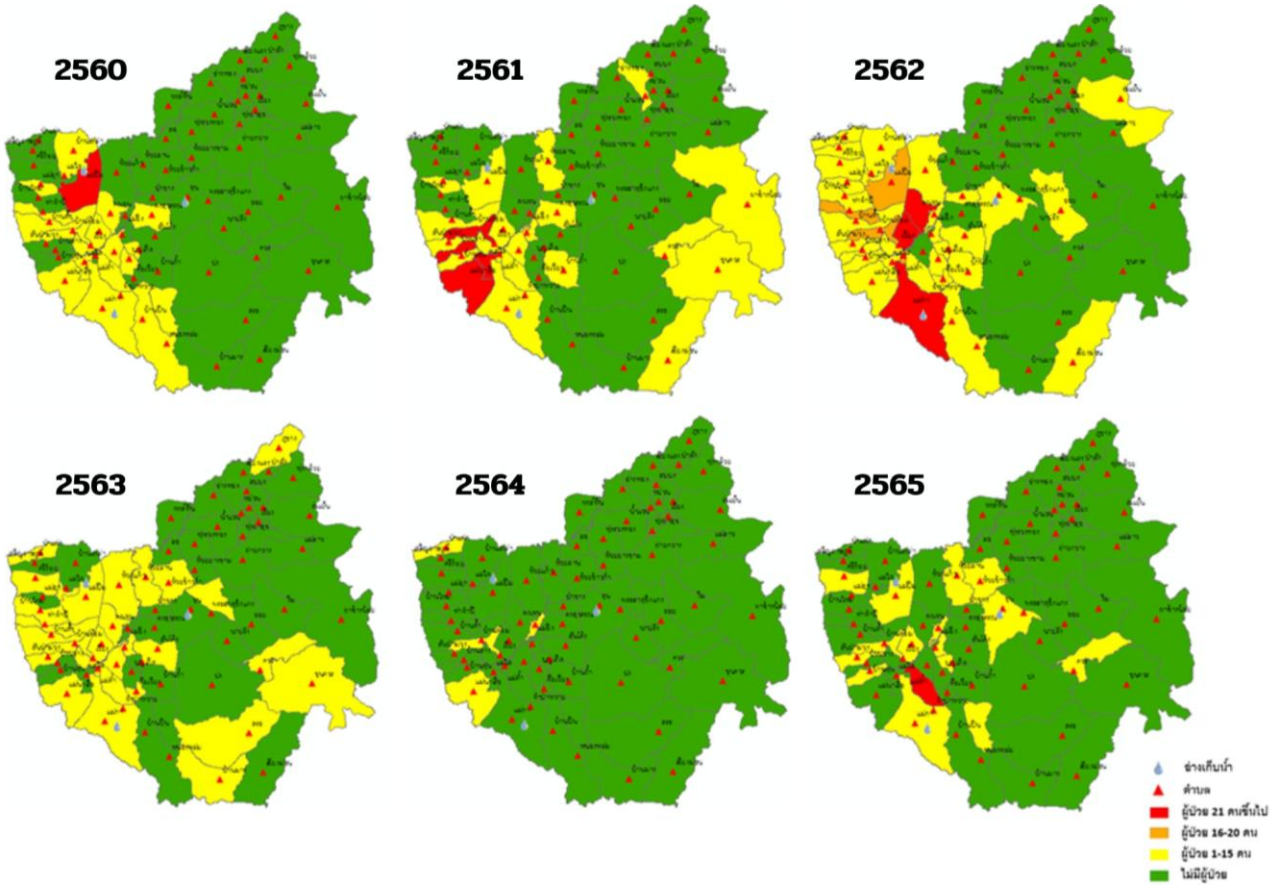
การศึกษาครั้งนี้เป็นการวิจัยเชิงวิเคราะห์ (Analytical Research) เพื่อศึกษาแนวทางการเรียนรู้ของเครื่องเพื่อการทำนายโรคไข้เลือดออก กรณีศึกษาจังหวัดพะเยา เพื่อทำนายการระบาดของจำนวนผู้ป่วยโรคไข้เลือดออกในจังหวัดพะเยา

แผนที่การระบาดของพื้นที่ศึกษา

การระบาดของโรคไข้เลือดออกในจังหวัดพะเยาโดยแบ่งเป็นรายอำเภอและรายตำบล ตั้งแต่ปี พ.ศ. 2560 – 2565 พบว่า ปี พ.ศ. 2562 พบผู้ป่วยไข้เลือดออกมากที่สุด จำนวน 262 คน โดยมีผู้ป่วยจำนวนมากในอำเภอเมืองพะเยา แม่ใจ ดอกคำใต้ ภูกามยาว เชียงม่วน จุน เชียงคำ ปง ตามลำดับ ดังรูปที่ 1 และปี พ.ศ. 2564 มีจำนวนผู้ป่วยน้อยที่สุด จำนวน 8 คน โดยมีผู้ป่วยในอำเภอเมืองพะเยา ภูกามยาว แม่ใจ ตามลำดับ ดังรูปที่ 1 หากพิจารณาในระดับตำบล ในปี พ.ศ. 2562 พบผู้ป่วยจำนวนมากที่สุด ในตำบลท่าวังทอง อำเภอเมืองพะเยา จำนวน 22 คน ผู้ป่วยน้อยที่สุดในตำบลเวียง สันป่าม่วง อำเภอเมืองพะเยา ตำบลละ 1 คน และพบผู้ป่วยน้อยที่สุดในปี พ.ศ. 2564 จำนวน 8 คน โดยมีตำบลแม่นาเรือ อำเภอเมือง 3 คน ตำบลบ้านต๋อม อำเภอเมือง 2 คน ตำบลแม่อิง อำเภอภูกามยาว 2 คน และ ตำบลป่าแฝก อำเภอแม่ใจ 1 คน ดังรูปที่ 2



รูปที่ 1 แผนที่การระบาดของโรคไข้เลือดออกในจังหวัดพะเยา ตั้งแต่ปี พ.ศ. 2560 – 2565 จำแนกเป็นรายอำเภอ



รูปที่ 2 แผนที่การระบาดของโรคไข้เลือดออกในจังหวัดพะเยา ตั้งแต่ปี พ.ศ. 2560 – 2565 จำแนกเป็นรายตำบล

การเก็บรวบรวมข้อมูล

ผู้วิจัยได้เก็บรวบรวมข้อมูลที่มา 3 แหล่ง โดยได้รับการอนุเคราะห์ข้อมูลจำนวนผู้ป่วยในโรงพยาบาลพะเยา ข้อมูลสภาพอากาศ จากสถานีอุตุนิยมวิทยาพะเยา และข้อมูลประชากรจากฐานข้อมูล สำนักงานสาธารณสุขจังหวัดพะเยา โดยจะแยกข้อมูลที่ใช้เป็นรายสัปดาห์และรายเดือน ซึ่งใช้ข้อมูลตั้งแต่ปี พ.ศ. 2560 – 2565 โดยมีรายละเอียด ดังตารางที่ 1

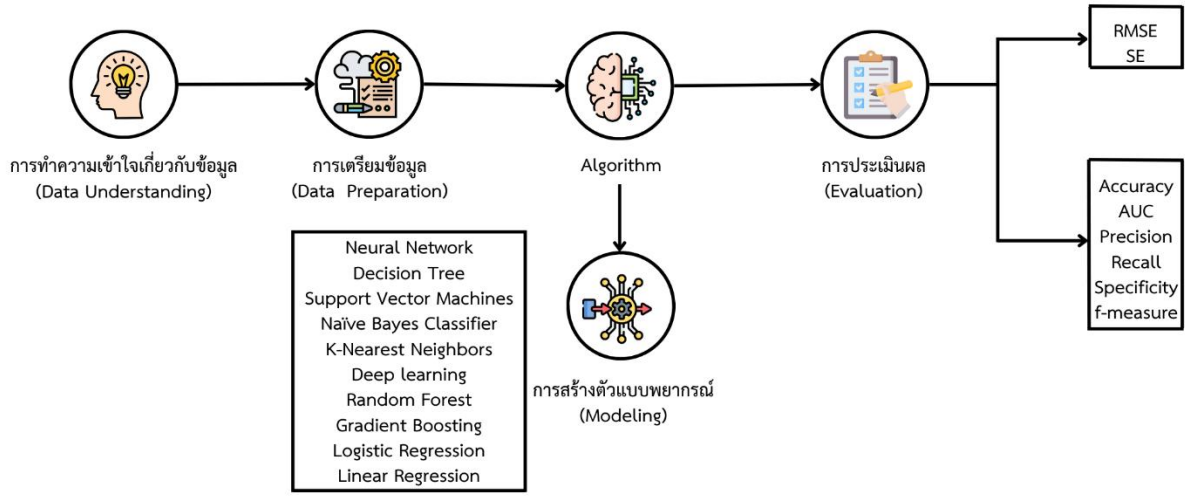
ตารางที่ 1 ข้อมูลที่ใช้ในการศึกษา

ข้อมูล	รายละเอียด	ปี	แหล่งที่มา
ข้อมูลจำนวนผู้ป่วยโรคไข้เลือดออก	รหัสของโรคและอาการ ICD10 A90, A91 วัน/เดือน/ปี ที่เข้ารับบริการ เพศ อายุ อำเภอ ตำบล หมู่ จำนวนผู้ป่วย	ปี 2560 – 2565	โรงพยาบาลพะเยา
ข้อมูลสภาพภูมิอากาศ	วัน/เดือน/ปี อุณหภูมิสูงสุด–ต่ำสุด อุณหภูมิเฉลี่ย ความชื้นสัมพัทธ์เฉลี่ย ปริมาณน้ำฝนเฉลี่ย จำนวนวันที่ฝนตก ในพื้นที่จังหวัดพะเยา	ปี 2560 – 2565	สถานีอุตุนิยมวิทยาพะเยา
ข้อมูลจำนวนประชากร	ปี พ.ศ. รหัสอำเภอ กลุ่มอายุ กลุ่มเพศ ประชากรรวม	ปี 2560 – 2565	สำนักงานสาธารณสุขจังหวัดพะเยา

การวิเคราะห์ข้อมูล

ผู้วิจัยได้วิเคราะห์ข้อมูลโดยใช้ข้อมูลจำนวนผู้ป่วยที่ได้จากโรงพยาบาลพะเยาและข้อมูลสภาพอากาศจากสถานีอุตุนิยมวิทยาพะเยา โดยแยกข้อมูลออกเป็นชุดข้อมูล (dataset) ซึ่งมีปัจจัยที่นำมาใช้ คือ อำเภอ ตำบล กลุ่มเพศ กลุ่มอายุ อุณหภูมิเฉลี่ย ความชื้นสัมพัทธ์เฉลี่ย ปริมาณน้ำฝนเฉลี่ย จำนวนวันที่ฝนตก และจำนวนประชากร ทำการแยกเป็นรายสัปดาห์ รายเดือน ตั้งแต่ปี พ.ศ. 2560 – 2565 ดังตารางที่ 2 เพื่อนำมาสร้างแบบจำลองการพยากรณ์ด้วยโปรแกรม RapidMiner (Mierswa & Klinkenberg, 2018)

ในการดำเนินงานวิจัยได้ใช้กระบวนการของ Data mining ดังรูปที่ 3 ด้วยการแปลงข้อมูลให้อยู่ในรูปแบบที่ต้องการ และสร้างแบบจำลองการพยากรณ์ เพื่อทำนายโรคไข้เลือดออก ด้วยเทคนิคการเรียนรู้ของเครื่อง ประกอบด้วยประเภท Regression ได้แก่ ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines) และการถดถอยเชิงเส้น (Linear Regression) ส่วนประเภท Classification ได้แก่ ต้นไม้ตัดสินใจ (Decision Tree) การถดถอยโลจิสติกส์ (Logistic Regression) โครงข่ายประสาทเทียม (Artificial Neural Network) นาอีฟเบย์ (Naïve Bayes) อัลกอริทึมการส่งเสริมประเภทหนึ่ง (Gradient Boosting Tree) ขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด (K–Nearest Neighbors) ต้นไม้สุ่ม (Random Forest) และ การเรียนรู้เชิงลึก (Deep Learning) โดยแยกข้อมูลที่วิเคราะห์เป็น Regression และ Classification จากนั้นใช้วิธี 5–fold Cross Validation เป็นการแบ่งข้อมูลเป็น 5 ส่วนเท่าๆ กัน โดยในแต่ละรอบจะใช้ 4 ส่วนสำหรับฝึกโมเดล และอีก 1 ส่วนสำหรับทดสอบ ทำซ้ำ 5 ครั้ง โดยในแต่ละครั้งจะเปลี่ยนส่วนที่ใช้ทดสอบ (Satangmongkol, K., 2019) เพื่อทำการวัดประสิทธิภาพและเปรียบเทียบประสิทธิภาพของแบบจำลองการพยากรณ์ ด้วยค่า Root Mean Square Error, Squared Error, Accuracy, AUC, Precision, Recall, Specificity และ F–measure ในส่วนของค่าพารามิเตอร์ที่กำหนดให้แต่ละแบบจำลองการพยากรณ์นั้นได้ใช้การปรับค่าพารามิเตอร์แบบอัตโนมัติโดยใช้เครื่องมือ Optimize Parameters ในโปรแกรม RapidMiner ซึ่งมีรายละเอียดกระบวนการ ดังรูปที่ 4



รูปที่ 3 กระบวนการสร้างแบบจำลองการพยากรณ์

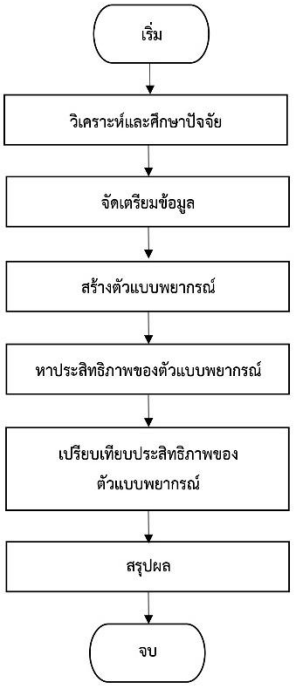
ตารางที่ 2 ชุดข้อมูลที่ใช้ในการศึกษา

ชุดข้อมูล	รายละเอียด	หมายเหตุ
ข้อมูลแยกตามอำเภอแบบรายเดือน (district–month)	อำเภอ เดือน ปี จำนวนประชากรแต่ละปี อุณหภูมิเฉลี่ยรายเดือน ปริมาณน้ำฝนเฉลี่ยรายเดือน ความชื้นสัมพัทธ์เฉลี่ยรายเดือน จำนวนวันที่ฝนตกต่อเดือน	(กรณีวิเคราะห์แบบ Regression จะใช้ Label จำนวนผู้ป่วยต่อประชากรแสนคนรายเดือน–รายสัปดาห์/กรณีวิเคราะห์แบบ Classification จะใช้ Label กลุ่มผู้ป่วยเป็น/ไม่เป็น)
ข้อมูลแยกตามอำเภอแบบรายสัปดาห์ (district–week)	อำเภอ สัปดาห์ ปี จำนวนประชากรแต่ละปี อุณหภูมิเฉลี่ยรายสัปดาห์ ปริมาณน้ำฝนเฉลี่ยรายสัปดาห์ ความชื้นสัมพัทธ์เฉลี่ยรายสัปดาห์ จำนวนวันที่ฝนตกต่อสัปดาห์	
ข้อมูลแยกตามตำบลแบบรายเดือน (subdistrict–month)	ตำบล เดือน ปี จำนวนประชากรแต่ละปี อุณหภูมิเฉลี่ยรายเดือน ปริมาณน้ำฝนเฉลี่ยรายเดือน ความชื้นสัมพัทธ์เฉลี่ยรายเดือน จำนวนวันที่ฝนตกต่อเดือน	
ข้อมูลแยกตามตำบลแบบรายสัปดาห์ (subdistrict–week)	ตำบล สัปดาห์ ปี จำนวนประชากรแต่ละปี อุณหภูมิเฉลี่ยรายสัปดาห์ ปริมาณน้ำฝนเฉลี่ยรายสัปดาห์ ความชื้นสัมพัทธ์เฉลี่ยรายสัปดาห์ จำนวนวันที่ฝนตกต่อสัปดาห์	
ข้อมูลแยกตามกลุ่มอายุแบบรายเดือน (age–month)	เดือน ปี จำนวนประชากรแต่ละปี อุณหภูมิเฉลี่ยรายเดือน ปริมาณน้ำฝนเฉลี่ยรายเดือน ความชื้นสัมพัทธ์เฉลี่ยรายเดือน จำนวนวันที่ฝนตกต่อเดือน กลุ่มอายุ	
ข้อมูลแยกตามกลุ่มอายุแบบรายสัปดาห์ (age–week)	สัปดาห์ ปี จำนวนประชากรแต่ละปี อุณหภูมิเฉลี่ยรายสัปดาห์ ปริมาณน้ำฝนเฉลี่ยรายสัปดาห์ ความชื้นสัมพัทธ์เฉลี่ยรายสัปดาห์ จำนวนวันที่ฝนตกต่อสัปดาห์ กลุ่มอายุ	
ข้อมูลแยกตามกลุ่มเพศแบบรายเดือน (sex–month)	เดือน ปี จำนวนประชากรแต่ละปี อุณหภูมิเฉลี่ยรายเดือน ปริมาณน้ำฝนเฉลี่ยรายเดือน ความชื้นสัมพัทธ์เฉลี่ยรายเดือน จำนวนวันที่ฝนตกต่อเดือน กลุ่มเพศ	
ข้อมูลแยกตามกลุ่มเพศแบบรายสัปดาห์ (sex–week)	สัปดาห์ ปี จำนวนประชากรแต่ละปี อุณหภูมิเฉลี่ยรายสัปดาห์ ปริมาณน้ำฝนเฉลี่ยรายสัปดาห์ ความชื้นสัมพัทธ์เฉลี่ยรายสัปดาห์ จำนวนวันที่ฝนตกต่อสัปดาห์ กลุ่มเพศ	

กระบวนการสร้างแบบจำลองการพยากรณ์

ขั้นตอนที่ 1 การวิเคราะห์ข้อมูล

การระบาดของโรคไข้เลือดออก อาจมีสาเหตุเนื่องมาจากหลายปัจจัย เช่น เพศ อายุ อุณหภูมิเฉลี่ย ปริมาณน้ำฝนเฉลี่ย ความชื้นเฉลี่ย และจำนวนวันที่ฝนตก ซึ่งเกี่ยวข้องกับเชื้อไวรัสและพาหะ ในปัจจุบันโรคไข้เลือดออกยังมีการแพร่ระบาดอย่างต่อเนื่องในชุมชน และเนื่องจากประชาชนไม่ได้มีวิธีแก้ปัญหาในพื้นที่หรือการป้องกันโรคไข้เลือดออกที่เพียงพอก่อนเกิดการระบาด ดังนั้นผู้วิจัยจึงทำการศึกษาโรคไข้เลือดออกโดยจะศึกษาปัจจัยที่มีผลต่อการระบาดของโรคไข้เลือดออก โดยจะใช้ข้อมูล ดังตารางที่ 3 และใช้เทคนิคการเรียนรู้ของเครื่องเพื่อสร้างแบบจำลองการพยากรณ์ ในการทำนายโรคไข้เลือดออกโดยให้ค่าการพยากรณ์ที่แม่นยำ เพื่อที่จะสามารถหาแนวทางป้องกันล่วงหน้าได้อย่างมีประสิทธิภาพ และช่วยลดโอกาสในการเกิดการระบาดของโรค



รูปที่ 4 ขั้นตอนการดำเนินงาน

ขั้นตอนที่ 2 การเตรียมข้อมูล

การเตรียมข้อมูลและเลือกปัจจัยสำคัญที่มีผลต่อการนำไปวิเคราะห์ข้อมูลเพื่อใช้ในการทำนายการระบาดโรคไข้เลือดออกจังหวัดพะเยา โดยใช้เทคนิคการเรียนรู้ของเครื่อง มีขั้นตอนดังต่อไปนี้

1. นำเข้าข้อมูลสำหรับการวิเคราะห์ข้อมูลดิบ โดยเก็บรวบรวมมาจากข้อมูลย้อนหลัง ตั้งแต่ปี พ.ศ. 2560 – 2565 ของโรงพยาบาลพะเยา สถานีอุตุนิยมวิทยาพะเยา สำนักงานสาธารณสุขจังหวัดพะเยา สำหรับการวิเคราะห์ข้อมูล

2. ตรวจสอบข้อมูลสูญหาย หากเป็นข้อมูลเชิงปริมาณจะแทนค่าด้วยค่าเฉลี่ยที่มีอยู่หรือข้อมูลปริมาณที่เกี่ยวข้อง หากเป็นข้อมูลเชิงคุณภาพจะแทนค่าด้วยค่าฐานนิยมแทนค่าข้อมูลที่สูญหาย เพื่อเตรียมข้อมูลให้พร้อมสำหรับการประมวลผล

3. กำหนดและตรวจสอบความถูกต้องของประเภทข้อมูล ตรวจสอบประเภทข้อมูลที่นำเข้าเพื่อให้แน่ใจว่าข้อมูลถูกต้องตามรูปแบบที่ต้องการในการวิเคราะห์และแก้ไขข้อมูลที่มีปัญหาหรือไม่สมบูรณ์ให้เรียบร้อย

ตารางที่ 3 ปัจจัยที่ใช้ในการศึกษา

ปัจจัย	คำอธิบาย	ประเภทข้อมูล	แหล่งที่มา
patient	จำนวนผู้ป่วยโรคไข้เลือดออก	จำนวนจริง	โรงพยาบาลพะเยา
age	กลุ่มอายุ	จำนวนจริง	
sex	กลุ่มเพศ	ทวิภาค	
district	อำเภอ	พหุนาม	
subdistrict	ตำบล	พหุนาม	
week	สัปดาห์	จำนวนเต็ม	
month	เดือน	พหุนาม	
year	ปี	จำนวนเต็ม	
population	จำนวนประชากร	จำนวนจริง	สำนักงานสาธารณสุขจังหวัดพะเยา
temperature	อุณหภูมิเฉลี่ยรายสัปดาห์/รายเดือน (เซลเซียส)	จำนวนจริง	สถานีอุตุนิยมวิทยาพะเยา
precipitation	ปริมาณน้ำฝนเฉลี่ยรายสัปดาห์/รายเดือน (มิลลิลิตร)	จำนวนจริง	
humidity	ความชื้นสัมพัทธ์เฉลี่ยรายสัปดาห์/รายเดือน (เปอร์เซ็นต์)	จำนวนจริง	
rain-day	จำนวนวันที่ฝนตกรายสัปดาห์/รายเดือน วัน)	จำนวนเต็ม	

ขั้นตอนที่ 3 การสร้างแบบจำลองการพยากรณ์

เป็นกระบวนการนำข้อมูลแต่ละปัจจัยที่เลือกมาเพื่อสร้างแบบจำลองการพยากรณ์ ด้วยเทคนิคการเรียนรู้ของเครื่องที่ผู้วิจัยได้คัดเลือกมาให้เหมาะสมสำหรับการทดสอบและจำแนกข้อมูลอย่างมีประสิทธิภาพที่สุด อัลกอริทึมที่ได้ทำการทดสอบ ได้แก่ Support Vector Machines (SVM), Linear Regression (LinR), Decision Tree (DT), Logistic Regression (LR), Artificial Neural Network (A-NN), Naïve Bayes (NB), Gradient Boosting Tree (GBT), K-Nearest Neighbors (K-NN), Random Forest (RF) และ Deep Learning (DL) โดยจะใช้การแบ่งข้อมูลในแต่ละแบบจำลองด้วยวิธี 5-Fold Cross Validation

ในส่วนของการปรับพารามิเตอร์ของแต่ละอัลกอริทึม สำหรับ Linear Regression ใช้ Max Iterations ในช่วง 1 ถึง 100 และค่า Bias ส่วน Support Vector Machines ใช้ Kernel Gramma โดยมีค่าต่ำสุดที่ 0 และสูงสุดที่ 100 และปรับค่าทีละ 10 บนสเกลเชิงลอการิทึม ส่วน Decision Tree และ Gradient Boosting Tree ใช้ Criterion โดยเลือก Information Gain, Accuracy, Gainratio, Gini Index และ Minimal Leaf Size กับ Max Depth เลือกอยู่ในช่วง 1 ถึง 100 โดยปรับค่าทีละ 10 บนสเกลเชิงเส้น ส่วน Logistic Regression ใช้ Standardize และ Lambda ในส่วนของ Lambda เลือกอยู่ในช่วง 1 ถึง 100 โดยปรับค่าทีละ 10 บนสเกลเชิงเส้น ส่วน Artificial Neural Network ใช้ Learning Rate ปรับค่าในช่วง 0.01 ถึง 1.0 โดยปรับค่าทีละ 10 บนสเกลเชิงเส้น และ ใช้ Hidden Layers จำนวน 5 ชั้น แต่ละชั้นใช้จำนวน 5 โหนด ส่วน Naïve Bayes ใช้ Laplace Correction ส่วน K-Nearest Neighbors ใช้จำนวน K ปรับค่าในช่วง 1 ถึง 100 โดยปรับค่าทีละ 10 บนสเกลเชิงเส้น Kernel Shift ปรับค่าในช่วง 1.0 ถึง 100 โดยปรับค่าทีละ 10 บนสเกลเชิงลอการิทึม และ Weighted vote ส่วน Random Forest ใช้ Criterion โดยเลือก Information Gain, accuracy, Gainratio, และ Gini Index และ Deep Learning ใช้ Activation ได้แก่ Tanh, TanhWithDropout, Rectifier, RectifierWithDropout, Maxout, MaxoutWithDropout, ExpRectifier และ ExpRectifierWithDropout ส่วน Epochs ปรับค่าในช่วง 1.0 ถึง 1.8 โดยปรับค่าทีละ 10 บนสเกลเชิงเส้น และใช้ Hidden layers จำนวน 2 ชั้น แต่ละชั้นใช้จำนวน 50 โหนด

ขั้นตอนที่ 4 การประเมินผล

ประเมินผลจากการวิเคราะห์ข้อมูลสำหรับการเรียนรู้และทำนายโรคไข้เลือดออก โดยเลือกข้อมูลสำหรับการเรียนรู้ข้อมูลที่มีประกอบด้วยปัจจัยต่าง ๆ เช่น จำนวนผู้ป่วย อายุ เพศ อุณหภูมิเฉลี่ย ความชื้นเฉลี่ย ปริมาณน้ำฝนเฉลี่ย และจำนวนวันที่ฝนตก โดยเลือกอัลกอริทึมที่จะทำการทดสอบและอธิบายผลการประเมินที่ได้จากการสร้างแบบจำลองพยากรณ์โรคไข้เลือดออก รายละเอียดของการวัดผลการประเมินประสิทธิภาพที่ใช้ (Ninenox Developer, 2020) มีดังนี้

1. Accuracy (ค่าความถูกต้อง) คือ ค่าที่บ่งบอกว่าแบบจำลอง ทำนายถูกต้องกี่เปอร์เซ็นต์จากทั้งหมด ควรมีค่าใกล้ *100%* หมายถึงทำนายถูกต้องมาก ดังสมการที่ (1)

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

2. Precision (ค่าความแม่นยำ) คือ ค่าที่บ่งบอกว่าเมื่อแบบจำลองทำนายว่าเป็น Positive มีความถูกต้องเท่าไร ควรมีค่าเข้าใกล้ 1 หมายถึงมีการทำนาย Positive ที่ถูกต้องมาก ดังสมการที่ (2)

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

3. AUC (ค่าการประเมินประสิทธิภาพ) คือ ค่าที่ใช้ในการประเมินประสิทธิภาพของแบบจำลองการจัดหมวดหมู่ ควรมีค่าระหว่าง 0 ถึง 1 โดยค่าเข้าใกล้ 1 หมายถึงประสิทธิภาพที่ดีที่สุด

4. Recall (ค่าความครบถ้วน) คือ ค่าที่บอกว่าแบบจำลองสามารถค้นหาค่าจริงที่เป็น Positive ได้ครบถ้วนเพียงใด ควรมีค่าสูงใกล้ 1 หมายถึงสามารถค้นหาค่าจริงที่เป็น Positive ได้ครบ ดังสมการที่ (3)

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

5. Specificity (ค่าความจำเพาะ) คือ ค่าที่บอกว่าแบบจำลองสามารถค้นหาค่าจริงที่เป็น Negative ได้ครบถ้วนเพียงใด ควรมีค่าสูงใกล้ 1 หมายถึงสามารถค้นหาค่าจริงที่เป็น Negative ได้ครบ ดังสมการที่ (4)

$$Specificity = \frac{TN}{TN+FP} \quad (4)$$

โดยที่

True Positive (TP) = จำนวนของตัวอย่างที่โมเดลทำนายว่าเป็น positive และค่าจริงเป็น positive หรือสามารถแยกแยะได้ถูกต้อง

False Positive (FP) = จำนวนของตัวอย่างที่โมเดลทำนายว่าเป็น positive แต่ค่าจริงไม่ใช่ positive หรือสามารถแยกแยะได้ผิดพลาด

True Negative (TN) = จำนวนของตัวอย่างที่โมเดลทำนายว่าเป็น negative และค่าจริงเป็น negative หรือสามารถแยกแยะได้ถูกต้อง

False Negative (FN) = จำนวนของตัวอย่างที่โมเดลทำนายว่าเป็น negative แต่ค่าจริงไม่ใช่ negative หรือสามารถแยกแยะได้ผิดพลาด

6. F–measure (ค่าความถ่วงดุล) คือ ค่าที่รวม Precision และ Recall เพื่อให้เห็นถึงประสิทธิภาพของแบบจำลองในมุมมองที่สมดุล ควรมีค่าสูงใกล้ 1 หมายถึงแบบจำลองมีความแม่นยำและครบถ้วน ดังสมการที่ (5)

$$F - measure = \frac{2 \times Percision \times Recall}{Percision + Recall} \quad (5)$$

7. Root Mean Squared Error (ค่าความคลาดเคลื่อนรวม) คือ ค่าที่บอกถึงความคลาดเคลื่อนของการทำนายที่เกิดขึ้นจริง โดยมีการถ่วงน้ำหนักจากค่า Error ที่สูงมากกว่า ควรมีค่าต่ำ หรือ มีค่าเข้าใกล้ 0 – 1 แสดงว่าการทำนายมีความคลาดเคลื่อนน้อย ดังสมการที่ (6)

$$RMSE = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(y_{i-}\hat{y}_i)^2} \quad (6)$$

โดยที่

n = จำนวนข้อมูลทั้งหมดในชุดข้อมูลที่ใช้ในการทดสอบ

y_i = ค่าจริงของข้อมูลในตำแหน่งที่ i

$\hat{y}_i$ = ค่าที่โมเดลทำนายในตำแหน่งที่ i

8. Squared Error (ค่าความคลาดเคลื่อนในรูปแบบพื้นที่) คือ ค่าคลาดเคลื่อนในรูปแบบพื้นที่ที่แสดงผลรวมของความคลาดเคลื่อนของแต่ละจุดข้อมูล ควรมีค่าที่ต่ำ ดังสมการที่ (7)

$$SE = (y - \hat{y})^2 \quad (7)$$

โดยที่

y = ค่าจริงของข้อมูล

$\hat{y}$ = ค่าที่โมเดลทำนาย

ขั้นตอนที่ 5 การนำไปใช้

หลังจากที่ได้สร้างแบบจำลองการพยากรณ์ ที่มีประสิทธิภาพแล้วด้วยเทคนิคการเรียนรู้ของเครื่องและทำการวัดประสิทธิภาพด้วยวิธีการตรวจสอบไขว้ (Cross – Validation) ขั้นตอนต่อไปคือการนำแบบจำลองการพยากรณ์นี้ไปใช้ เพื่อการใช้งานจริงในสถานการณ์ที่ต้องการทำนายหรือจำแนกประเภทของข้อมูลต่าง ๆ ขั้นตอนการนำไปใช้ของแบบจำลองการพยากรณ์นั้นมีหลายวิธี โดยจะขึ้นอยู่กับลักษณะของการนำไปใช้และแพลตฟอร์มที่ต้องการใช้งาน

ผลการศึกษา

ผลการวิเคราะห์ข้อมูลผู้ป่วยโรคไข้เลือดออกพบว่า จำนวนผู้ป่วยโรคไข้เลือดออกจากโรงพยาบาลปี พ.ศ. 2560 – 2565 มีแนวโน้มเพิ่มขึ้นจากปี พ.ศ. 2560 จำนวน 68 ราย ปี พ.ศ. 2561 จำนวน 78 ราย ปี พ.ศ. 2562 จำนวน 262 ราย ปี พ.ศ. 2563 จำนวน 246 ราย ส่วนปี พ.ศ. 2564 มีผู้ป่วยลดลงจำนวน 8 รายและปี พ.ศ. 2565 มีผู้ป่วยเพิ่มขึ้นจำนวน 121 ราย ซึ่งสัมพันธ์กับสภาพอากาศที่ได้จากสถานีอุตุนิยมวิทยาพะเยาและได้นำมาสร้าง

แบบจำลอง 2 ประเภท คือ Regression และ Classification โดยนำข้อมูลมาสร้าง dataset จำนวน 8 datasets ได้แก่ ข้อมูลแยกตามอำเภอแบบรายเดือน (district–month), ข้อมูลแยกตามอำเภอแบบรายสัปดาห์ (district–week), ข้อมูลแยกตามตำบลแบบรายเดือน (subdistrict–month), ข้อมูลแยกตามตำบลแบบรายสัปดาห์ (subdistrict–week), ข้อมูลแยกตามกลุ่มอายุแบบรายเดือน (age–month), ข้อมูลแยกตามกลุ่มอายุแบบรายสัปดาห์ (age–week), ข้อมูลแยกตามกลุ่มเพศแบบรายเดือน (sex–month) และ ข้อมูลแยกตามกลุ่มเพศแบบรายสัปดาห์ (sex–week) เพื่อเปรียบเทียบประสิทธิภาพของแต่ละแบบจำลองการพยากรณ์

ผลการวัดประสิทธิภาพของแบบจำลองการพยากรณ์ ประเภท Regression ทั้ง 8 datasets ได้แก่ district–month, district –week, subdistrict–month, subdistrict–week, age–month, age–week, sex–month และ sex–week โดยใช้วิธีการ Linear Regression ซึ่งให้ค่า RMSE ที่ 9.697, 3.511, 73.132, 30.794, 5.557, 2.100, 2.815 และ 1.190 ตามลำดับ เมื่อเทียบกับผลที่ได้จากการใช้วิธีการ Support Vector Machines ซึ่งให้ค่า RMSE ที่ 10.337, 8.558, 77.368, 31.291, 5.921, 8.127, 3.097 และ 6.166 ตามลำดับ ตามตารางที่ 4 พบว่า วิธีการ Linear Regression ให้ประสิทธิภาพดีกว่าวิธีการ Support Vector Machines ในทุก datasets ทั้ง 8 แบบ

ตารางที่ 4 ผลการวัดประสิทธิภาพของแบบจำลองการพยากรณ์ ประเภท Regression

ชุดข้อมูล (dataset)	**Support Vector Machine**		**Linear Regression**	
	RMSE	**SE**	**RMSE**	**SE**
district–month	10.337	111.185	9.697	94.031
district –week	8.558	73.330	3.511	12.327
subdistrict–month	77.368	6564.470	73.132	53478.270
subdistrict–week	31.291	982.351	30.794	948.289
age–month	5.921	37.412	5.557	31.554
age–week	8.127	66.162	2.100	4.438
sex–month	3.097	11.146	2.815	7.925
sex–week	6.166	38.058	1.190	1.454

ผลการวัดประสิทธิภาพของแบบจำลองการพยากรณ์ ประเภท Classification แบบรายเดือน ใน dataset district–month พบว่า Deep Learning ให้ประสิทธิภาพดีที่สุด ตามด้วย Decision Tree, Gradient Boosting Tree, Random Forest, Artificial Neural Network, Logistic Regression, K–Nearest Neighbors และ Naive Bayes ที่ Accuracy 95.31, 92.11, 91.00, 88.42, 87.22, 83.43, 74.50 และ 70.56 ตามลำดับ subdistrict–month พบว่า Deep Learning ให้ประสิทธิภาพดีที่สุด ตามด้วย Gradient Boosting Tree, Decision Tree, Logistic Regression, Random Forest, Artificial Neural Network, Naive Bayes และ K–Nearest Neighbors ที่ Accuracy 99.74, 98.98, 97.22, 93.29, 92.92, 86.48, 86.07 และ 74.95 ตามลำดับ age–month พบว่า Deep Learning ให้ประสิทธิภาพดีที่สุด ตามด้วย Decision Tree, Gradient Boosting Tree, Random Forest, Logistic Regression, Artificial Neural Network, Naive Bayes และ K–Nearest Neighbors ที่ Accuracy 96.69, 89.01, 85.71, 85.25, 82.24, 82.08, 78.62 และ 71.43 ตามลำดับ sex–month พบว่า Deep learning ให้ประสิทธิภาพดีที่สุด ตามด้วย Decision Tree, Gradient Boosting Tree, Logistic Regression, Naive Bayes, Random Forest, Artificial Neural Network และ K–Nearest Neighbors ที่ Accuracy 88.94, 86.90, 85.00, 81.37, 80.94, 78.33, 72.44 และ 70.00 ตามลำดับ

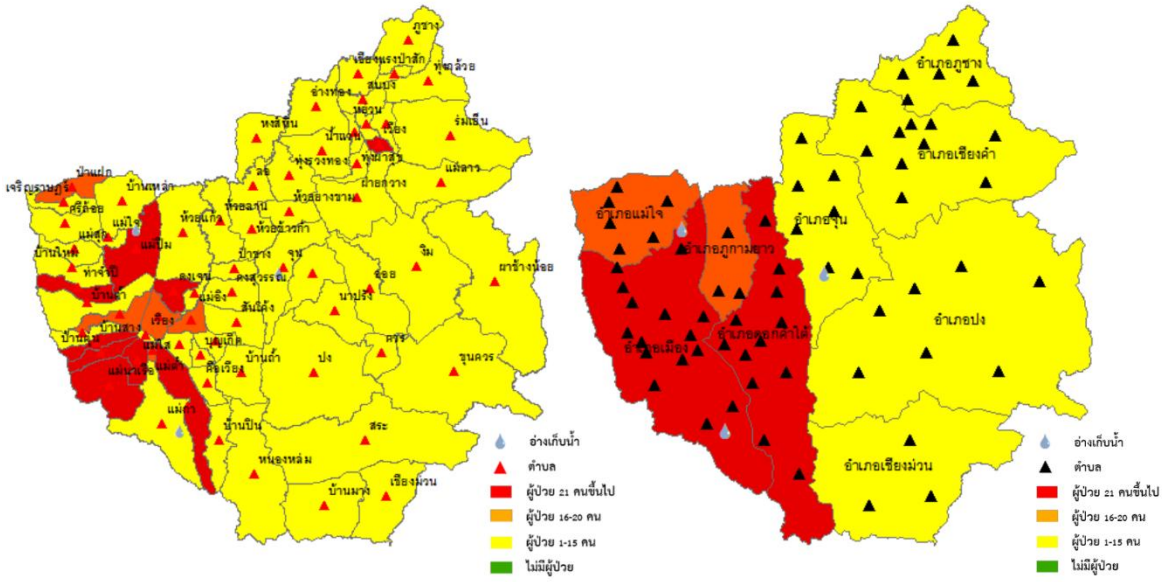
ส่วนผลการวัดประสิทธิภาพ แบบรายสัปดาห์ ใน dataset district–week พบว่า Deep Learning ให้ประสิทธิภาพดีที่สุด ตามด้วย Gradient Boosting Tree, Decision Tree, Random Forest, Logistic Regression, Artificial Neural Network, K–Nearest Neighbors และ Naive Bayes ที่ Accuracy 99.45, 97.13, 95.22, 92.53, 87.06, 86.07, 77.06 และ 77.05 ตามลำดับ subdistrict–week พบว่า Deep Learning ให้ประสิทธิภาพดีที่สุด ตามด้วย Gradient Boosting Tree, K–Nearest Neighbors, Decision Tree, Random Forest, Logistic Regression, Artificial Neural Network และ Naive Bayes ที่ Accuracy 99.84, 99.62, 99.16, 97.62, 96.84, 96.06, 87.05 และ 69.48 ตามลำดับ age–week พบว่า Deep Learning ให้ประสิทธิภาพดีที่สุด ตามด้วย Decision Tree, Gradient Boosting Tree, Random Forest, Artificial Neural Network, Logistic Regression, Naive Bayes และ K–Nearest Neighbors ที่ Accuracy 96.93, 90.33, 88.80, 84.39, 77.63, 77.28, 73.19 และ 68.17 ตามลำดับ sex–week พบว่า Deep Learning ให้ประสิทธิภาพดีที่สุด ตามด้วย Gradient Boosting Tree, Decision Tree, Random Forest, Artificial Neural Network, Logistic Regression, Naive Bayes และ K–Nearest Neighbors ที่ Accuracy 91.79, 85.96, 85.11, 77.14, 75.14, 74.79, 71.28 และ 59.06 ตามลำดับ ตามตารางที่ 5

จากผลประยุกต์ใช้ตัวแบบพยากรณ์ ผู้วิจัยได้นำตัวแบบที่ให้ประสิทธิภาพดีที่สุดมาทำนายจำนวนผู้ป่วยไข้เลือดออกปี 2566 ของจังหวัดพะเยา ซึ่งผลของการทำนายได้นำเสนอในรูปแบบของแผนที่ แสดงดังรูปที่ 5

ตารางที่ 5 ผลการวัดประสิทธิภาพของแบบจำลองการพยากรณ์ ประเภท Classification

Model	district–month						district–week					
	Accuracy	AUC	Precision	Recall	Specificity	f–measure	Accuracy	AUC	Precision	Recall	Specificity	f–measure
DT	92.11	0.950	99.10	85.73	99.19	91.90	95.22	0.976	91.44	99.73	90.75	95.41
LR	83.43	0.904	86.50	81.20	86.31	83.69	87.06	0.928	82.81	93.51	80.62	87.82
A–NN	87.22	0.880	90.02	85.28	89.48	87.55	86.07	0.904	82.81	90.95	81.19	86.67
NB	70.56	0.874	95.66	46.38	97.69	62.44	77.05	0.599	69.30	96.88	57.33	80.78
GBT	91.00	0.953	92.86	89.22	92.86	91.00	97.13	0.988	0.988	99.25	94.90	97.26
K–NN	74.50	0.819	70.97	85.44	62.89	77.53	77.06	0.874	73.45	81.12	73.36	77.10
RF	88.42	0.898	95.79	81.66	95.9	88.14	92.53	0.954	87.62	99.04	86.03	92.97
DL	95.31	0.979	100.00	91.21	100.00	95.36	99.45	0.996	98.90	100.00	98.90	99.45
Model	subdistrict–month						subdistrict–week					
	Accuracy	AUC	Precision	Recall	Specificity	f–measure	Accuracy	AUC	Precision	Recall	Specificity	f–measure
DT	97.22	0.963	94.74	100.00	91.70	97.30	97.62	0.757	52.63	4.05	99.91	7.75
LR	93.29	0.95	89.20	98.58	88.00	93.65	96.06	0.85	22.17	28.80	97.73	24.66
A–NN	86.48	0.884	81.29	95.00	77.92	87.58	87.05	0.891	81.99	95.36	78.78	88.11
NB	86.07	0.500	78.31	100.00	72.00	87.83	69.48	0.500	5.72	76.20	69.32	10.64
GBT	98.98	0.992	97.98	100.00	97.99	98.98	99.62	0.998	99.48	99.76	99.49	99.62
K–NN	74.95	0.829	77.78	69.32	80.51	73.35	99.16	0.500	98.34	100.00	98.32	99.16
RF	92.92	0.949	88.81	98.42	87.39	93.34	96.84	0.982	94.34	99.75	93.96	96.95
DL	99.74	0.998	99.49	100.00	99.48	99.75	99.84	0.999	99.73	99.96	99.73	99.84

Model	age-month						age-week					
	Accuracy	AUC	Precision	Recall	Specificity	f-measure	Accuracy	AUC	Precision	Recall	Specificity	f-measure
DT	89.01	0.855	88.75	96.32	73.02	92.36	90.33	0.942	84.13	99.05	81.25	91.10
LR	82.24	0.867	84.64	90.26	64.68	87.30	77.28	0.863	73.03	87.15	66.18	79.3
A-NN	82.08	0.826	84.83	89.90	65.38	87.23	77.63	0.815	76.32	80.23	75.13	78.16
NB	78.62	0.831	77.77	96.23	39.76	85.98	73.19	0.692	66.75	92.37	53.89	77.49
GBT	85.71	0.860	88.66	91.49	71.79	90.05	88.80	0.962	90.98	86.38	91.27	88.62
K-NN	71.43	0.659	71.43	100	0.00	83.33	68.17	0.734	64.92	78.26	58.20	70.97
RF	85.25	0.778	85.06	95.13	64.12	89.76	84.39	0.877	79.33	93.37	75.68	85.65
DL	96.69	0.956	95.92	99.36	91.06	97.60	96.93	0.982	94.92	99.22	94.65	97.01
Model	sex-month						sex-week					
	Accuracy	AUC	Precision	Recall	Specificity	f-measure	Accuracy	AUC	Precision	Recall	Specificity	f-measure
DT	86.90	0.964	100.00	71.10	100.00	82.87	85.11	0.917	77.72	86.82	73.70	81.86
LR	81.37	0.84	80.84	76.18	85.27	78.33	74.79	0.83	72.33	81.45	70.40	76.55
A-NN	72.44	0.804	82.69	50.18	92.81	60.99	75.14	0.812	73.71	79.47	70.61	76.48
NB	80.94	0.818	81.22	73.27	85.80	76.68	71.28	0.721	66.45	87.77	54.12	75.62
GBT	85.00	0.907	75.00	85.71	84.62	80.00	85.96	0.923	82.61	90.48	81.61	86.36
K-NN	70.00	0.786	64.29	56.25	79.17	60.00	59.06	0.636	56.60	71.43	47.13	63.16
RF	78.33	0.708	77.32	71.17	83.78	73.95	77.14	0.801	74.41	84.33	69.61	78.98
DL	88.94	0.895	100.00	75.71	100.00	85.98	91.79	0.939	91.79	92.15	91.17	91.96



รูปที่ 5 แผนที่แสดงผลการพยากรณ์การเกิดโรคไข้เลือดออก ปี พ.ศ. 2566 รายตำบล และรายอำเภอ

วิจารณ์และสรุปผล

การวิจัยในครั้งนี้ผู้วิจัยได้ใช้วิธีการเลือกโดยการใช้ปัจจัยที่มีความสัมพันธ์กับตัวแปรกลุ่มเป้าหมายแล้วนำมาสร้างแบบจำลองการพยากรณ์ สำหรับการทำนายการระบาดของโรคไข้เลือดออกในจังหวัดพะเยา ได้แก่ ปัจจัยด้านพื้นที่ คือ จังหวัด อำเภอ และตำบล โดยการระบาดของโรคไข้เลือดออกมีความแตกต่างกันในแต่ละพื้นที่ เนื่องจากความหนาแน่นของประชากร สภาพแวดล้อม และการกระจายตัวของแหล่งเพาะพันธุ์ยุงลาย ซึ่งเป็นพาหะของโรคไข้เลือดออก พื้นที่ที่มีแหล่งน้ำขังที่สะสมไว้เป็นเวลานานจะมีอัตราการระบาดสูงกว่า ส่วนปัจจัยด้านเวลา คือ ปีที่ตรวจพบเชื้อไข้เลือดออก อาจจะมีแนวโน้มเปลี่ยนแปลงตามช่วงเวลา เช่น ความถี่ของการระบาดอาจจะเพิ่มขึ้นในบางปี เนื่องจากวงจรการระบาดที่เกิดจากการหมุนเวียนของสายพันธุ์ไวรัสหรือการเปลี่ยนแปลงในประชากรยุงลายที่เป็นพาหะ และปัจจัยด้านสภาพภูมิอากาศ คือ อุณหภูมิเฉลี่ย โดยอุณหภูมิที่สูงขึ้นจะเพิ่มอัตราการเจริญเติบโตและการแพร่พันธุ์ของยุงลาย อีกทั้งยังส่งผลต่อความสามารถในการแพร่เชื้อไวรัสในยุง ส่วนปริมาณน้ำฝนเฉลี่ยและจำนวนวันที่ฝนตก โดยฝนที่ตกทำให้เกิดแหล่งน้ำขังที่เหมาะสมสำหรับการวางไข่และการเจริญเติบโตของยุงลาย ยิ่งฝนตกมากเท่าไร โอกาสที่ยุงลายจะเพิ่มขึ้นก็ยิ่งมาก และความชื้นสัมพัทธ์เฉลี่ย โดยความชื้นที่สูงส่งผลให้ยุงลายมีอายุยืนขึ้น และเพิ่มโอกาสในการแพร่กระจายเชื้อไข้เลือดออก โดยปัจจัยดังกล่าวใช้เป็นรายเดือนและรายสัปดาห์ และได้สอดคล้องกับปัจจัยที่ใช้ในการศึกษาการพยากรณ์การระบาดโรคไข้เลือดออกในจังหวัดพะเยา พบว่า ปัจจัยด้านพื้นที่ ปัจจัยด้านเวลาและปัจจัยสภาพภูมิอากาศเป็นปัจจัยที่มีผลในการเกิดการระบาดโรคไข้เลือดออก และผู้วิจัยได้พบว่า ผลการวัดประสิทธิภาพของแบบจำลองการพยากรณ์ ประเภท Regression พบว่า Linear Regression ให้ประสิทธิภาพดีที่สุด ที่ RMSE 1.190 ใน dataset sex-week ซึ่งมีค่าความคลาดเคลื่อนที่ต่ำบ่งชี้ว่าแบบจำลองนี้มีความแม่นยำสูงในการทำนาย ส่วนผลที่ได้ประเภท Classification พบว่า Deep Learning ให้ประสิทธิภาพดีที่สุด ที่ Accuracy 99.84% ใน dataset subdistrict-week ซึ่งเป็นแบบจำลองสำหรับการพยากรณ์การระบาดโรคไข้เลือดออกที่เหมาะสมที่สุด โดยให้ค่าความแม่นยำและค่าประสิทธิภาพโดยรวมสูงที่สุด ซึ่งสอดคล้องกับงานวิจัยแบบจำลองการพยากรณ์การระบาดของโรคไข้เลือดออกโดยใช้เทคนิคการทำเหมืองข้อมูลพบว่าวิธี Decision Tree เป็นโมเดลในการสร้างแบบจำลองสำหรับการพยากรณ์การระบาดของโรคไข้เลือดออกที่เหมาะสมที่สุดโดยให้ค่าความแม่นยำสูงสุด 69.83% และ 75.4% ตามลำดับ (จิรโรจน์ ตอสะสุกุล, 2564) และในงานวิจัยการทำนายไข้เลือดออกในสิงคโปร์ มีการใช้แนวทางการเรียนรู้ของเครื่องที่ โดยใช้ข้อมูลอุตุนิยมวิทยาในการสร้างแบบจำลอง พบว่าการสร้างแบบจำลองด้วยเทคนิค XGBoost แสดงประสิทธิภาพที่ดีที่สุด โดยมีค่า MAE = 89.12, RMSE = 156.07 และ R-squared = 0.83 (Tian et al., 2024) และงานวิจัยการพยากรณ์กรณีไข้เลือดออกในรัฐคุชราตโดยใช้ Long Short-Term Memory แนวทางการเรียนรู้ของเครื่อง พบว่า LSTM ให้ค่าความแม่นยำที่ดีที่สุด 84% ที่ RMSE อยู่ที่ 0.04 และค่าคะแนน R-squared อยู่ที่ 0.84 (Mehta & Patel, 2023) และงานวิจัยแนวทางการเรียนรู้เชิงลึกเพื่อทำนายไข้เลือดออก ได้ใช้ Deep Learning โดยใช้โมเดล LSTM ในการทำนายผลโดยแบบจำลองที่ได้ผลดีที่สุดคือ SSA-LSTM อยู่ที่ค่า RMSE ที่ 3.17 แบบจำลอง SSA-LSTM ทำงานได้ดีในขอบเขตการทำนายของงานวิจัยนี้ (Majeed et al, 2023) ซึ่งข้อมูลที่ใช้ในงานวิจัยดังกล่าวนั้นมีความแตกต่างกับงานวิจัยนี้ คือ ข้อมูลกลุ่มอาการของโรคไข้เลือดออกกับอาชีพของผู้ที่ติดเชื้อ ความเร็วลม ความกดอากาศที่ระดับน้ำทะเล รังสีจากดวงอาทิตย์ พลังงานจากดวงอาทิตย์ ดัชนีรังสียูวี

จากผลการศึกษา พบว่า Deep Learning มีประสิทธิภาพดีกว่าโมเดลอื่น ๆ อาจเนื่องจาก อัลกอริทึม สามารถสร้างและเรียนรู้คุณลักษณะเชิงซ้อนจากข้อมูลได้ดีกว่า ซึ่งจะเป็นประโยชน์อย่างมากในการวางแผนและจัดการสาธารณสุข โดยเฉพาะในพื้นที่ที่มีความเสี่ยงสูง โดยโครงข่ายประสาทเทียมใน Deep Learning มีความสามารถในการทำความเข้าใจข้อมูลที่มีโครงสร้างซับซ้อนและมีความสัมพันธ์ที่ไม่เป็นเชิงเส้นในข้อมูลได้ดี อีกทั้งการใช้จำนวนชั้น (layers)

ที่มากกว่าในโครงข่ายประสาทเทียมทำให้ Deep Learning สามารถสร้างแบบจำลองในรูปแบบที่ซับซ้อนได้มากขึ้น ทำให้หน่วยงานที่เกี่ยวข้องสามารถทำนายการระบาดได้แม่นยำมากยิ่งขึ้น และมีการปรับแต่งและเพิ่ม ประสิทธิภาพจากกระบวนการเรียนรู้ที่หลากหลาย ทำให้สามารถจัดการกับข้อมูลที่มีปริมาณมากและหลากหลายได้อย่างมีประสิทธิภาพ

ข้อจำกัดของงานวิจัยนี้คือ ข้อมูลที่ใช้ในการวิเคราะห์อาจไม่สมบูรณ์ และความแตกต่างของพื้นที่ในการวิเคราะห์ความแตกต่างในแง่ของการระบาด ซึ่งอาจมีความคลาดเคลื่อนในสถานการณ์จริง สำหรับในงานวิจัยครั้งต่อไปควรศึกษาเพิ่มเติมเกี่ยวกับปัจจัยที่อาจมีผลต่อการเกิดโรคไข้เลือดออก หรือปัจจัยที่มีความสัมพันธ์กับกลุ่มเป้าหมาย เช่น อาชีพของผู้ติดเชื้อ หรือปัจจัยทางสิ่งแวดล้อมเพิ่มเติม เพื่อเพิ่มความแม่นยำในการพยากรณ์และช่วยในการวางแผนเชิงนโยบายที่มีประสิทธิภาพมากขึ้นในการควบคุมโรคไข้เลือดออก

กิตติกรรมประกาศ

ผู้วิจัยขอขอบคุณโรงพยาบาลพะเยา สถานีอุตุนิยมวิทยาพะเยา และสำนักงานสาธารณสุขจังหวัดพะเยา ที่ให้ความอนุเคราะห์ข้อมูล และให้คำปรึกษาและข้อเสนอแนะในการทำวิจัยครั้งนี้

เอกสารอ้างอิง

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

กลุ่มโรคติดต่อระหว่างประเทศ กองโรคติดต่อทั่วไป กรมควบคุมโรค. (2563). *ไข้เลือดออก (Dengue Fever และ Dengue Hemorrhagic Fever) (ICD 10 A90, A91)*. พะเยา: กรมควบคุมโรค.

กาญจนา ยังขาว และ กัญญรัตน์ สระแก้ว. (2558*)*. การพยากรณ์โรคไข้เลือดออกโดยใช้ข้อมูล 5 มิติเขตสุขภาพที่ 9. *วารสารควบคุมโรค, 41*(3), 208–218.

กอบเกียรติ สระอุบล. (2563). *เรียนรู้ Data Science และ AI: Machine Learning ด้วย Python*. กรุงเทพฯ: หสม มีเดีย เนทเวิร์ค.

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

ชาญชัยณรงค์ ทรงคาศรี. (2555). รูปแบบการพยากรณ์โรคไข้เลือดออกในพื้นที่สำนักงานป้องกันควบคุมโรคที่ 6 จังหวัดขอนแก่น พ.ศ. 2555. *วารสารวิชาการสำนักป้องกันควบคุมโรคที่ 7 ขอนแก่น, 20*(1), 65–81.

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

สำนักโรคติดต่อนำโดยแมลง สำนักระบาดวิทยา. (2560). *รายงานพยากรณ์โรค "ไข้เลือดออก" ปี 2560* (ฉบับที่ 1 ปี 2560). พะเยา.

สายชล สินสมบูรณ์ทอง. (2560). *การทำเหมืองข้อมูล เล่ม 1: การค้นหาความรู้จากข้อมูล*. กรุงเทพฯ: บริษัทจามจุรี โปรดักส์ จำกัด.

สุรศักดิ์ สุขสาย. (2550). การพยากรณ์พื้นที่เสี่ยงต่อการเกิดโรคไข้เลือดออกเพื่อการวางแผนเฝ้าระวังและป้องกันในจังหวัดอุบลราชธานี. *วารสารวิจัย มข. (บศ.), 7*(2), 83–96. สืบค้นจาก https://aigencorp.com/what-is-machine-learning-technology.

Chaw, J. K., Chaw, S. H., Quah, C. H., Sahrani, S., Ang, M. C., Zhao, Y., & Ting, T. T. (2023). A predictive analytics model using machine learning algorithms to estimate the risk of shock development among dengue patients, *Healthcare Analytics, 5*, 2–7.

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), 2–4. Form: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10002017/.

Mehta, A. M., & Patel, K. S. (2023), LSTM–based Forecasting of Dengue Cases in Gujarat: A Machine Learning Approach, *Indian Journal of Science and Technology, 17*(7), 635 – 642. Form: https://indjst.org/articles/lstm–based–forecasting–of–dengue–cases–in–gujarat–a–machine–learning–approach.

Mierswa, I., & Klinkenberg, R. (2018). *RapidMiner Studio (9.1) [Data science, machine learning, predictive analytics]*. Form: https://rapidminer.com/.

Ninenox Developer. (2020). Understand accuracy, precision, recall, f1–score. Form: https://www.ninenox.com/2020/09/24/–accuracyprecisionrecallf1–score/.

Satangmongkol, K.. (2019). *Explaining K–Fold Cross Validation with sample code in R.* Form: https://datarockie–com.translate.goog/blog/k–fold–cross–validation/comment–page–.

Sebastianelli, A., Spiller, D., Carmo, R., Wheeler, J., Nowakowski, A., Jacobson, V. A., Kim, D., Barlevi, H., Cordero, Z. E. R., Colón–González, F. J., Lowe, R., Ullo, S. L., & Schneider, R. (2024). A reproducible ensemble machine learning approach to forecast dengue outbreaks. *Scientific Reports, 14*(3807). Form: https://www.nature.com/articles/s41598–024–52796–9.

Tian, N., Zheng, J. X., Li, L. H., Xue, J. B., Xia, S., Lv, S., & Zhou, X. N. (2024). Precision Prediction for Dengue Fever in Singapore: A Machine Learning Approach Incorporating Meteorological Data. *Tropical Medicine and Infectious Disease, 4*(9), 5–7. Form: https://pubmed.ncbi.nlm.nih.gov/38668533/.