การพยากรณ์โรคคล้ายไข้หวัดใหญ่ในประเทศไทย

ธรรศกรณ์ เศวตสุทธิพันธ์[*1], นราภรณ์ เภาประเสริฐ[2] และ พาพิศ วงศ์ชัยสุวัฒน์[3]
ภาควิชาวิศวกรรมอุตสาหการ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ กรุงเทพมหานคร 10900

Received: 25 February 2024; Revised: 02 May 2024; Accepted: 23 May 2024

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อวิเคราะห์ข้อมูลการระบาดของโรคไข้หวัดใหญ่ในประเทศไทย โดยการศึกษาปัจจัยที่เกี่ยวข้องกับโรคไข้หวัดใหญ่และทำการพยากรณ์อัตราการติดเชื้อโรคคล้ายไข้หวัดใหญ่ (Influenza-like illness percentage : ILI%) ซึ่งค่า ILI% ถูกรวบรวมในรูปแบบรายเดือนของแต่ละจังหวัดในประเทศไทย โดยเปรียบเทียบการพยากรณ์ทั้ง 5 วิธี ได้แก่ สมการถดถอยเชิงเส้นพหุคูณ, Regression tree, Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost) และRandom forest โดยคุณลักษณะที่นำมาใช้ประกอบด้วยปัจจัยด้านวัคซีน ปัจจัยด้านโรคกลุ่มเสี่ยง ปัจจัยด้านประชากร รวมถึงปัจจัยด้านสภาพอากาศ อีกทั้งยังมีการใช้วิธีในการคัดเลือกคุณลักษณะ (Feature Selection) เพื่อหาความสัมพันธ์ระหว่างคุณลักษณะ และค่าตัวแปรตอบสนอง ได้แก่ วิธี Stepwise, การจัดลำดับความสำคัญด้วยค่า Features Importance, การจัดลำดับความสำคัญด้วยค่า SHAP, วิธี Boruta, วิธี BorutaSHAP และ การใช้ค่า Mutual Information Scores (MI Scores) ร่วมกับ Boruta และ BorutaSHAP ในการประเมินประสิทธิภาพของแบบจำลอง ผู้วิจัยได้ใช้ค่าคลาดเคลื่อนเฉลี่ยร้อยละแบบสมมาตร (SMAPE) เป็นตัวชี้วัด ผลลัพธ์ที่ได้คือ วิธี Random forest ที่มีการคัดเลือกคุณลักษณะโดยใช้ค่า MI Scores ร่วมกับ BorutaSHAP ให้ค่า SMAPE ต่ำสุดในชุดข้อมูลทดสอบอยู่ที่ 59.11% และมีคุณลักษณะที่สำคัญ ได้แก่ อัตราการฉีดวัคซีน, จำนวนบ้าน, จำนวนประชากรในกลุ่มอายุ 7-9 ปี, จำนวนประชากรในกลุ่มอายุ 15-24 ปี และจำนวนผู้ป่วยหลอดเลือดสมองตีบ ซึ่งจากทั้งหมดนี้การพยากรณ์ค่า ILI% เพื่อป้องกันและลดผลกระทบที่เกิดจากโรคระบาด และใช้ประกอบการตัดสินใจในการกระจายวัคซีน รวมถึงกลยุทธ์การป้องกันการแพร่ระบาดในระดับชุมชน

คำสำคัญ: การพยากรณ์โรคไข้หวัดใหญ่, โรคระบาด, ILI%, การเรียนรู้ของเครื่อง, การคัดเลือกคุณลักษณะ

* Corresponding author. E-mail: thassakorn.saw@ku.th
[1] นิสิต หลักสูตรดุษฎีบัณฑิต ภาควิชาวิศวกรรมอุตสาหการ คณะวิศวกรรมศาสตร์มหาวิทยาลัยเกษตรศาสตร์ (บางเขน)
[2] รองศาสตราจารย์ ภาควิชาวิศวกรรมอุตสาหการ คณะวิศวกรรมศาสตร์มหาวิทยาลัยเกษตรศาสตร์ (บางเขน)
[3] รองศาสตราจารย์ ภาควิชาวิศวกรรมอุตสาหการ คณะวิศวกรรมศาสตร์มหาวิทยาลัยเกษตรศาสตร์ (บางเขน)

Prediction of Influenza-Like Illness in Thailand

Thassakorn Sawetsuthipan[*1], Naraphorn Paoprasert[2] and Papis Wongchaisuwat[3]

Department of Industrial Engineering, Faculty of Engineering, Kasetsart University, Bangkok 10900

Received: 25 February 2024; Revised: 02 May 2024; Accepted: 23 May 2024

Abstract

This research aims to analyze the outbreak of influenza in Thailand by studying factors related to the epidemic and predicting the Influenza-like Illness percentage (ILI%). The ILI% data, aggregated monthly for each province in Thailand, is compared using five prediction methods: multiple linear regression, regression tree, Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), and random forest. The features used include vaccine-related factors, risk group disease factors, population factors, and weather factors. Additionally, feature selection methods such as Stepwise, Features Importance Ranking, SHAP Ranking, Boruta, BorutaSHAP, and Mutual Information Scores (MI Scores) with Boruta and BorutaSHAP. To evaluate the performance of the models, the researchers used the symmetric mean percentage error (SMAPE) as a metric. Random forest method, using MI Scores with BorutaSHAP, achieved the lowest SMAPE of 59.11% on the test dataset and identified significant features such as vaccination rate, number of houses, population aged 7-9 years, population aged 15-24 years, and number of patients with stroke. These forecasts can help prevent and mitigate the impact of outbreaks and inform vaccine distribution decisions, as well as community-level outbreak prevention strategies.

Keywords: influenza forecasting, epidemic, ILI%, machine learning, feature selection

* Corresponding author. E-mail: thassakorn.saw@ku.th

[1]Doctoral Student in Department of Industrial Engineering, Faculty of Engineering, Kasetsart University (Bangkhen)

[2]Associate Professor in Department of Industrial Engineering, Faculty of Engineering, Kasetsart University (Bangkhen)

[3]Associate Professor in Department of Industrial Engineering, Faculty of Engineering, Kasetsart University (Bangkhen)

1. บทนำ

ไข้หวัดใหญ่เป็นโรคระบาดทางเดินหายใจที่มีผลกระทบต่อสุขภาพของมนุษย์มาเป็นเวลานาน เกิดจากการติดเชื้อไวรัสไข้หวัดใหญ่ (Influenza virus) เกิดได้ตลอดทั้งปี โดยมีอุบัติการณ์สูงสุดในช่วงฤดูฝน และ ฤดูหนาว โรคไข้หวัดใหญ่ในคน ประกอบด้วยชนิด A ได้แก่ A (H1N1), A (H1N2), A (H3N2) และ A (H5N1) ชนิด B และชนิด C พบว่าชนิด A มีการแพร่ระบาดมากที่สุด คนเป็นแหล่งรังโรคไข้หวัดใหญ่ที่สำคัญ ทุกเพศทุกวัยมีโอกาสป่วย โดยเฉพาะผู้ป่วยโรคเรื้อรัง การป่วยเป็นโรคไข้หวัดใหญ่มีโอกาสเกิดภาวะแทรกซ้อนที่รุนแรงและเสียชีวิตได้ ข้อมูลเฝ้าระวังโรค กองระบาดวิทยา กรมควบคุมโรค กระทรวงสาธารณสุข ปี พ.ศ. 2565 ประเทศไทย มีรายงานผู้ป่วยโรคไข้หวัดใหญ่ 77,831 ราย คิดเป็นอัตราป่วย 117.62 คนต่อประชากรแสนคน กลุ่มอายุที่มีอัตราป่วยมากที่สุด คือ 0-4 ปี อัตราป่วย 540.06 คนต่อประชากร แสนคน รองลงมา คือ กลุ่มอายุ 5-9 ปี อัตราป่วย 403.15 คนต่อประชากรแสนคน [1] การพยากรณ์โรคไข้หวัดใหญ่จึงเป็นสิ่งสำคัญในการจัดการกับการระบาดของโรคและในการเตรียมความพร้อมในกรณีที่มีการระบาดมากขึ้น

การพยากรณ์โรคไข้หวัดใหญ่มีประโยชน์ในการช่วยกำหนดมาตรการป้องกันและควบคุมโรคที่เหมาะสมเพราะโรคไข้หวัดใหญ่มีความสามารถในการแพร่กระจายสูงและปรับตัวตามสภาวะสภาพแวดล้อมขณะนั้น ดังนั้นการทำการตรวจวินิจฉัยจากห้องปฏิบัติจึงเป็นเรื่องยากถ้าวิเคราะห์ผู้ป่วยแต่ละรายว่าเป็นโรคไข้หวัดใหญ่หรือไม่ ดังนั้น อาการโรคที่คล้ายกับการเป็นไข้หวัดใหญ่ (Influenza-like illness : ILI) มักถูกนำมาใช้เป็นตัวแทนในการวิเคราะห์โรคไข้หวัดใหญ่ [2] และการวิเคราะห์ข้อมูลเกี่ยวกับสภาวะสภาพแวดล้อม เช่น อุณหภูมิและความชื้นในอากาศ ก็เป็นสิ่งที่สำคัญในการพยากรณ์การระบาดของไข้หวัดใหญ่ การเข้าใจและการรับรู้ถึงปัจจัยเหล่านี้เป็นสิ่งสำคัญในการจัดการและควบคุมโรคไข้หวัดใหญ่เพื่อเตรียมความพร้อมและลดผลกระทบต่อสังคมและสุขภาพของมนุษย์

ในงานวิจัยนี้มุ่งเน้นพัฒนาและเปรียบเทียบตัวแบบพยากรณ์ที่แตกต่างกัน และศึกษาปัจจัยที่อาจส่งผลต่ออัตราการติดเชื้อโรคคล้ายไข้หวัดใหญ่ โดยขอบเขตงานวิจัยนี้ได้รวบรวมชุดข้อมูลจากกรมควบคุมโรค กระทรวงสาธารณสุขแห่งประเทศไทย ซึ่งใช้อัตราการติดเชื้อโรคคล้ายไข้หวัดใหญ่หรือค่า ILI% เป็นตัวแปรตอบสนอง (Response) และข้อมูลทางด้านประชากรศาสตร์ ลักษณะภูมิอากาศ และโรคกลุ่มเสี่ยง ถูกนำมาใช้เป็นคุณลักษณะ (Features) ผู้วิจัยได้ทำการเปรียบเทียบตัวแบบพยากรณ์ทั้ง 5 แบบได้แก่ สมการถดถอยเชิงเส้นพหุคูณ (Multiple linear regression : MLR), Regression tree, LightGBM, XGBoost และ Random forest ด้วยการออกแบบและเพิ่มเติมคุณลักษณะเข้าไป และใช้การคัดเลือกคุณลักษณะด้วยวิธีต่าง ๆ เช่น วิธี Stepwise, การคัดเลือกผ่านค่า Feature Importance, การคัดเลือกผ่านค่า SHAP (Shapley additive explanations), การคัดเลือกโดยใช้วิธี Boruta และ BorutaSHAP และการผสมระหว่างวิธี Boruta และ BorutaSHAP ร่วมกับค่า MI Scores และวัดผลโดยใช้ค่า SMAPE (Symmetric Mean Absolute Percentage Error)

2. การทบทวนวรรณกรรม

จากงานวิจัยก่อนหน้าในการพยากรณ์โรคไข้หวัดใหญ่สามารถแบ่งออกเป็น 3 สาขาหลักได้คือ โมเดลสถิติดั้งเดิม อัลกอริทึมการเรียนรู้ของเครื่อง และการเรียนรู้เชิงลึก ในการวิจัยโดยใช้โมเดลสถิติดั้งเดิม อย่างเช่น Sanguanrungsirikul et al. [3] ได้เปรียบเทียบวิธีพยากรณ์จำนวนผู้ติดเชื้อโรคไข้หวัดใหญ่ วิธีเฉลี่ยเคลื่อนที่แบบง่าย (Simple Moving Average method) วิธีปรับให้เรียบเอ็กซ์โพเนนเชียลแบบง่าย (Simple Exponential Smoothing method) วิธีสัดส่วนกับแนวโน้ม (Ratio-To-Trend method) วิธีการปรับให้เรียบแบบเอ็กซ์โพเนนเชียลแบบโฮลท์-วินเทอร์ (Exponential Smoothing Holt-Winter method) วิธีบอกซ์เจนกินส์ (Box-Jenkins method) และงานวิจัยอื่น ๆ ที่ใช้วิธีทางด้าน Autoregressive (AR) [4], Autoregressive Integrated Moving Average (ARIMA) [5-6] และ Seasonal Autoregressive Integrated Moving Average (SARIMA) [7] ซึ่งข้อจำกัดของการใช้โมเดลสถิติดั้งเดิมเหล่านี้ คือไม่ได้นำปัจจัยอื่น ๆ มาพิจารณาโดยทำการศึกษาข้อมูลอันเนื่องมาจากเวลาเท่านั้น ดังนั้นความสามารถในการทำนายของโมเดลสถิติดั้งเดิมเหล่านี้อาจไม่มีประสิทธิภาพเท่าที่ควร

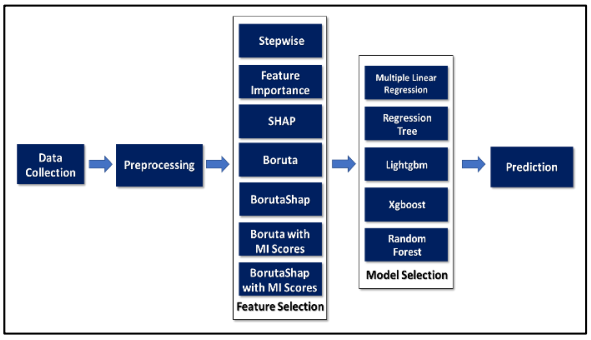
การเพิ่มประสิทธิภาพในการพยากรณ์โรคไข้หวัดใหญ่นักวิจัยและสถาบันหลายแห่งได้พัฒนาแบบจำลองต่าง ๆ ให้ดีขึ้นโดยแบบจำลองเหล่านั้นใช้วิธีการเรียนรู้ของเครื่อง เช่น การวิเคราะห์สมการถดถอยเชิงเส้นพหุคูณ [8-9] ซึ่งวิธีนี้มีปัญหาบางประการคือ ค่าของ ILI และปัจจัยที่เกี่ยวข้องนั้นอาจไม่ได้แสดงความสัมพันธ์ในเชิงเส้นตรง ดังนั้นการใช้สมการเชิงเส้นอาจไม่ได้ผลลัพธ์ที่มีประสิทธิภาพในการทำนายที่ต้องการ Xue et al. [10] ได้เปรียบเทียบการพยากรณ์ค่า ILI โดยใช้การวิเคราะห์สมการถดถอยหลายตัวแปร และแบบจำลองโครงข่ายประสาทเทียม ซึ่งแบบจำลองโครงข่ายประสาทเทียมถูกใช้ในการวิเคราะห์การสมการถดถอยแบบไม่เชิงเส้นโดยให้ผลลัพธ์ทำนายค่า ILI ที่ดีกว่า นอกจากนี้ Xue et al. [11] ได้เสนอวิธีการ Support vector regression (SVR) ที่ใช้ขั้นตอนการปรับปรุงอัลกอริทึม Particle swarm optimization (IPSO-SVR) ซึ่งถูกนำไปใช้ในโมเดลการทำนายไข้หวัดเพื่อทำนาย ILI% ในระดับภูมิภาคทางใต้ของประเทศจีน ซึ่งมีความแม่นยำและผลลัพธ์ที่ดีกว่าวิธี Multiple linear regression, SVR, GA-SVR, และ PSO-SVR แต่ยังมีข้อจำกัดในบางประการเช่น ไม่ได้พิจารณาปัจจัยโรคที่เกี่ยวข้องอื่น ๆ และมีขอบเขตเพียงแค่ทางภูมิภาคทางใต้ของจีนเท่านั้น ในการทำนายในภูมิภาคอื่น ๆ ยังไม่ได้รับการศึกษาอย่างเพียงพอ

ในส่วนของการเรียนรู้เชิงลึกนั้น Zhang & Nawata [12] ได้เปรียบเทียบการพยากรณ์แบบจำลองทั้ง 6 รูปแบบ ได้แก่ ARIMA, SVR , Random forest, Gradient boosting, Artificial neural network และ Long Short-Term Memory (LSTM) ในการพยากรณ์ค่า ILI พบว่า LSTM ให้ประสิทธิภาพในด้านความแม่นยำมากที่สุด แต่ยังมีข้อจำกัดในด้านการใช้ปัจจัยภายนอกมาร่วมใช้ในการพยากรณ์ เช่น สภาพอากาศ อุณภูมิและความชื้น และปัจจัยที่ไม่ควรมองข้ามอย่าง ขนาดประชากร ความหนาแน่นของประชากร และวิถีชีวิต เช่นเดียวกับ Kara [13] ได้ปรับปรุงประสิทธิภาพของแบบจำลอง LSTM ด้วยการใช้ genetic algorithm มาผสมโดยได้เปรียบเทียบการพยากรณ์แบบจำลองทั้ง 5 รูปแบบได้แก่ GA-LSTM, ARIMA, Random forest regression (RAF), support vector regression (SVR), และ fully-connected neural network (FCNN) ผลลัพธ์ยังคงพบว่าแบบจำลองการเรียนรู้เชิงลึกยังคงให้ผลลัพธ์ดีกว่า และยังคงข้อจำกัดในเรื่องของการไม่ได้ใช้ปัจจัยภายนอกต่าง ๆ เข้ามาพิจารณา เป็นเพียงการพยากรณ์ด้วยค่าตัวแปรตอบสนองเพียงอย่างเดียว

ในงานวิจัยที่กล่าวมานั้น วิธีการพยากรณ์จะมุ่งเน้นไปในเรื่องของพยากรณ์ด้วยอนุกรมเวลา โดยการพยากรณ์ด้วยวิธีการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกนั้นยังไม่ได้ศึกษาปัจจัยที่เกี่ยวข้องกับโรคไข้หวัดใหญ่เท่าที่ควร งานวิจัยของผู้วิจัยนั้นมุ่งเน้นไปที่การศึกษาปัจจัยซึ่งเป็นส่วนหนึ่งที่สำคัญสำหรับการพยากรณ์โรคไข้หวัดใหญ่ จากแนวคิดของ Stoppiglia et al. [14] วิธี Boruta เป็นเทคนิคที่ใช้ในอัลกอริทึม Random forest เพื่อการระบุตัวแปรนำเข้าที่เกี่ยวข้องโดยการเปรียบเทียบความสำคัญของตัวแปรที่แท้จริงกับตัวแปรสุ่ม ตัวแปรหลักที่ใช้จะถูกกำหนดโดยการกระจายของ metrics ของ Z-score เทคนิคนี้ลบตัวแปรซ้ำ ๆ ที่มีความสำคัญน้อยกว่าตัวแปรสุ่ม และวิธี SHAP (Shapley additive explanations) เป็นหนึ่งในวิธีที่ใช้เพื่อหาความสัมพันธ์ระหว่างตัวแปรนำเข้าและผลลัพธ์ของแบบจำลองโดยการคำนวณค่า SHAP สำหรับแต่ละตัวแปรนำเข้า [15] ค่า SHAP ในแบบจำลองเรียนรู้เครื่องแสดงถึงความสำคัญของตัวแปรนำเข้าโดยการเปรียบเทียบค่าทำนายสำหรับตัวอย่างนำเข้ากับค่าพยากรณ์เมื่อไม่มีตัวแปรนำเข้า โดยการวิเคราะห์โมเดลโดยใช้ SHAP สามารถกำหนดความสำคัญของตัวแปรนำเข้าและอธิบายว่าการพยากรณ์นั้นเปลี่ยนแปลงอย่างไรเมื่อค่าของตัวแปรนำเข้านั้นเปลี่ยนแปลง

3. วิธีการดำเนินงานวิจัย

วัตถุประสงค์ของการศึกษานี้คือการพยากรณ์ค่าอัตราการติดเชื้อโรคคล้ายไข้หวัดใหญ่ หรือ ILI% โดยใช้ข้อมูลของจังหวัดต่าง ๆ ในประเทศไทย รูปที่ 1 แสดงขั้นตอนการเก็บข้อมูล การจัดเตรียมข้อมูล รวมถึงวิธีการที่ใช้ในการพยากรณ์ ค่า ILI% ถูกใช้เป็นตัวแปรตอบสนอง ในส่วนของตัวแปรนำเข้านั้น ใช้ปัจจัยต่าง ๆ ได้แก่ ปัจจัยด้านสภาพอากาศ ปัจจัยทางด้านประชากรศาสตร์ อีกทั้งยังมีปัจจัยทางด้านโรคกลุ่มเสี่ยง และ จำนวนผู้ฉีดวัคซีน โดยการวิเคราะห์ข้อมูลทั้งหมดและการสร้างแบบจำลองพยากรณ์ถูกออกแบบบน Python 3.11 (64-bit)



รูปที่ 1 ขั้นตอนการดำเนินงาน

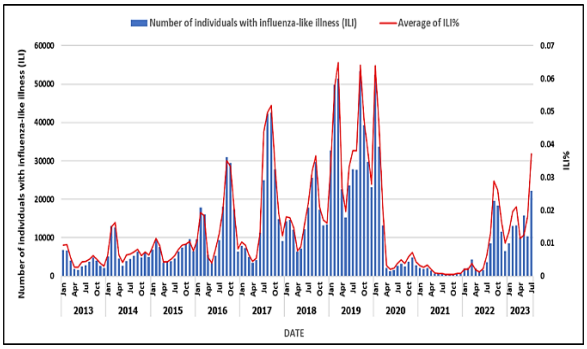
3.1 การรวบรวมข้อมูลและการเตรียมข้อมูล

ชุดข้อมูลที่ใช้ในการวิเคราะห์ค่า ILI% ถูกรวบรวมมาในรูปแบบรายเดือนของแต่ละจังหวัดตั้งแต่เดือนมกราคม พ.ศ. 2556 ถึง เดือน กรกฎาคม พ.ศ. 2566 ได้รวบรวมข้อมูลที่เกี่ยวข้องกับโรคไข้หวัดใหญ่ โรคกลุ่มเสี่ยงและวัคซีน ซึ่งได้รับมาจากกรมควบคุมโรคภายใต้กระทรวงสาธารณสุข ข้อมูลทางด้านสภาพอากาศจากกรมอุตุนิยมวิทยาประเทศไทยโดยในการวัดอุณหภูมิอากาศ (°F) และ ความชื้นสัมพัทธ์ (%) ถูกรวบรวมมาในรูปแบบค่าสถิติสูงสุด ต่ำสุด และค่าเฉลี่ย ใน ในส่วนข้อมูลทางด้านประชากรศาสตร์ ได้แก่ จำนวนประชากรในกลุ่มช่วงอายุต่าง ๆ ของแต่ละจังหวัดในประเทศไทย ถูกรวบรวมจากสำนักบริหารการทะเบียน กรมการปกครอง และสำหรับค่า ILI% ที่จะถูกใช้เป็นตัวแปรตอบสนองสำหรับการพยากรณ์นั้นสามารถคำนวณได้จากสมการที่ 1

$$ILI\% = \frac{\text{จำนวนผู้มีอาการคล้ายติดเชื้อโรคไข้หวัดใหญ่ } (ILI) \times 100}{\text{จำนวนประชากร } (N)} \quad (1)$$

เมื่อ ILI% คือ อัตราการติดเชื้อโรคคล้ายไข้หวัดใหญ่ โดยจำนวนประชากรที่ศึกษาจะใช้จำนวนประชากรในแต่ละจังหวัดในการได้มาซึ่งค่า ILI% ของแต่ละจังหวัด

ในชุดข้อมูลที่ได้รับมานั้นมีบางส่วนที่ไม่สมบูรณ์ เช่น การขาดหายไปของข้อมูล หรือค่าที่ผิดปกติ เพื่อแก้ไขปัญหาเหล่านี้จะใช้เทคนิคการทำความสะอาดข้อมูล (Data cleaning) โดยการแทนค่าเฉลี่ยของเดือนก่อนหน้าและเดือนถัดไปสำหรับข้อมูลที่ขาดหายไป รูปที่ 2 แสดงจำนวนผู้มีอาการคล้ายติดเชื้อโรคไข้หวัดใหญ่และค่าเฉลี่ย ILI% รายเดือนในประเทศไทย จากข้อมูล ผู้วิจัยพบว่าข้อมูลในช่วงเดือนเมษายน พ.ศ. 2563 ถึง เดือนมิถุนายน พ.ศ. 2565 ได้มีการระบาดของโรค Covid-19 จึงมีผลทำให้ข้อมูลมีค่าที่ผิดปกติ ในข้อมูลส่วนนี้จะไม่ถูกนำมาใช้ในการพยากรณ์ และในบางจังหวัดที่ไม่มีข้อมูลที่เกี่ยวข้องกับโรคกลุ่มเสี่ยงและวัคซีนจะถูกตัดออก ซึ่งจะเหลือข้อมูลทั้งหมด 5,800 จุด คุณลักษณะต่าง ๆ จะถูกทำให้อยู่ในรูปแบบ Min-Max Scaling (Normalization) ซึ่งเป็นการปรับค่าข้อมูลให้อยู่ในช่วง [0,1] เพื่อลดความแปรปรวนของข้อมูล



รูปที่ 2 จำนวนผู้ติดเชื้อโรคคล้ายไข้หวัดใหญ่รายเดือนในประเทศไทย

เพื่อทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปร ผู้วิจัยได้เพิ่มตัวแปรในรูปแบบเชิงประเภท (categorical data) ได้แก่ เดือนและภูมิภาคของจังหวัดนั้น ๆ เข้ามาวิเคราะห์เป็นคุณลักษณะเพิ่มเติมในการทดลอง ซึ่งคุณลักษณะที่รวบรวมมาและคุณลักษณะเพิ่มเติมจากตัวแปรในรูปแบบเชิงประเภททั้งหมดที่ใช้รวมทั้งสิ้น 63 คุณลักษณะ และคำอธิบายคุณลักษณะ แสดงตามตารางที่ 1

ตารางที่ 1 คุณลักษณะทั้งหมดที่ใช้ในการทดลอง

คุณลักษณะ	คำอธิบาย (หน่วย)	คุณลักษณะ	คำอธิบาย (หน่วย)
Age0-2	จำนวนประชากรตามกลุ่มอายุ (คน)	Wet Temp	อุณหภูมิตุ้มเปียก (องศาเซลเซียส)
Age3-6		Population density	ความหนาแน่นประชากร (คน/ตารางกิโลเมตร)
Age7-9		Population per household	จำนวนประชากรต่อบ้าน (คน/หลัง)
Age10-14		Vaccination_m	จำนวนผู้ฉีดวัคซีน (คน)
Age15-24		Vaccination rate	จำนวนผู้ฉีดวัคซีน (%)
Age25-34		Diabetes_I_m	ผู้ป่วยเบาหวาน (คน)
Age35-44		Diabetes rate	ผู้ป่วยเบาหวาน (%)
Age45-54		COPD_I_m	ผู้ป่วยหลอดลมอุดกั้นเรื้อรัง (คน)
Age55-64		COPD rate	ผู้ป่วยหลอดลมอุดกั้นเรื้อรัง (%)
Age65+		Stroke_I_m	ผู้ป่วยหลอดเลือดสมองตีบ (คน)
Age0-2 (%)	จำนวนประชากรตามกลุ่มอายุ (%)	Stroke rate	ผู้ป่วยหลอดเลือดสมองตีบ (%)
Age3-6 (%)		Cardiovascular_I_m	ผู้ป่วยหัวใจและหลอดเลือด (คน)
Age7-9 (%)		Cardiovascular rate	ผู้ป่วยหัวใจและหลอดเลือด (%)
Age10-14 (%)		Month1	มกราคม
Age15-24 (%)		Month2	กุมภาพันธ์
Age25-34 (%)		Month3	มีนาคม
Age35-44 (%)		Month4	เมษายน
Age45-54 (%)		Month5	พฤษภาคม
Age55-64 (%)		Month6	มิถุนายน
Age65+ (%)		Month7	กรกฎาคม
Total population	ประชากรทั้งหมด (คน)	Month8	สิงหาคม
Number of houses	จำนวนบ้าน (หลัง)	Month9	กันยายน
In-migration rate	อัตราการย้ายเข้า (%)	Month10	ตุลาคม
Out-migration rate	อัตราการย้ายออก (%)	Month11	พฤศจิกายน
Area	พื้นที่จังหวัด (ตารางกิโลเมตร)	Month12	ธันวาคม
Max Humidity	ค่าความชื้นสัมพัทธ์ (%)	Region_C	ภาคกลาง
Avg Humidity		Region_E	ภาคตะวันออก
Min Humidity		Region_N	ภาคเหลือ
Max Temp	อุณหภูมิ (องศาเซลเซียส)	Region_NE	ภาคตะวันออกเฉียงเหนือ
Avg Temp		Region_S	ภาคใต้
Min Temp		Region_W	ภาคตะวันตก
rain	ปริมาณฝน (มิลลิเมตร)		

3.2 การสร้างแบบจำลองพยากรณ์

เพื่อสร้างแบบจำลองการเรียนรู้ของเครื่อง ในงานวิจัยนี้ ผู้วิจัยได้แบ่งข้อมูลแบบสุ่ม โดยแบ่งข้อมูลออกเป็น 80% ซึ่งจะถูกใช้เป็นชุดข้อมูลฝึกสอน (Train set) สำหรับการสร้างแบบจำลอง และ 20% จะถูกใช้เป็นชุดข้อมูลทดสอบ (Test set) สำหรับประเมินประสิทธิภาพ โดยรูปแบบการพยากรณ์ที่ใช้จะอยู่ในรูปของสมการถดถอย (Regression Models) หรือการใช้คุณลักษณะในการพยากรณ์ตัวแปรตอบสนอง ในการสร้างแบบจำลองเบื้องต้นทุกแบบจำลองจะใช้คุณลักษณะทุกตัวในการสร้าง (Full Model) และแต่ละแบบจำลองจะใช้เทคนิคการเลือกคุณลักษณะที่แตกต่างกัน โดยแบบจำลองและเทคนิคการเลือกคุณลักษณะที่นำมาใช้ได้แก่

1) สมการถดถอยเชิงเส้นพหุคูณ เป็นวิธีการทางสถิติที่ใช้ในการวิเคราะห์และพยากรณ์ข้อมูลที่มีตัวแปรต้นมากกว่าหนึ่งตัวแปรและนำมาสร้างสมการถดถอยเชิงเส้น

การคัดเลือกคุณลักษณะของแบบจำลองนี้จะใช้วิธี Stepwise ซึ่งเป็นวิธีการคัดเลือกตัวแปรต้นที่มีความสำคัญกับตัวแปรตามให้มากที่สุดและลดตัวแปรที่ไม่สำคัญออก

2) Regression tree เป็นแบบจำลองที่สร้างจากการแบ่งข้อมูลออกเป็นกลุ่มย่อย ๆ โดยใช้เงื่อนไขการแบ่งข้อมูลที่มีค่าสัมประสิทธิ์สูงที่สุด เหมาะกับข้อมูลที่มีความซับซ้อนและมีตัวแปรต้นหลายตัวแปร การคัดเลือกคุณลักษณะของแบบจำลองนี้จะใช้วิธีการจัดลำดับความสำคัญของคุณลักษณะโดยใช้ค่า Features Importance ซึ่งเป็นค่าทางสถิติโดยจะประเมินว่าคุณลักษณะมีผลกระทบกับค่าตัวแปรตอบสนองเท่าใด โดยทางผู้วิจัยจะคัดเลือกคุณลักษณะที่มีค่า Features Importance 15 อันดับที่มีค่ามากสุดในการนำมาสร้างแบบจำลอง

3) LightGBM เป็นอัลกอริทึมที่ใช้ในการสร้างแบบจำลองเชิงเรียนรู้ถูกพัฒนาขึ้นจากต้นไม้ตัดสินใจ (Decision Tree) เพื่อใช้วิเคราะห์และพยากรณ์ข้อมูลที่มีขนาดใหญ่และมีความซับซ้อน โดยเฉพาะข้อมูลที่มีจำนวนคุณลักษณะมาก ในการคัดเลือกคุณลักษณะจะใช้วิธีการจัดลำดับความสำคัญของคุณลักษณะโดยใช้ค่า Features Importance เช่นเดียวกับแบบจำลอง Regression tree

4) XGBoost เป็นอัลกอริทึมที่ใช้ในการสร้างแบบจำลองเชิงเรียนรู้ถูกพัฒนาขึ้นจากต้นไม้ตัดสินใจ เพื่อใช้วิเคราะห์และพยากรณ์ข้อมูลที่มีจำนวนคุณลักษณะมากที่มีกระจายตัวและความสัมพันธ์ซับซ้อนระหว่างคุณลักษณะ การคัดเลือกคุณลักษณะจะใช้วิธีการจัดลำดับความสำคัญของคุณลักษณะโดยใช้ค่า Features Importance และ ค่า SHAP ในการกำหนดคุณลักษณะ 15 อันดับแรก

5) Random forest เป็นอัลกอริทึมที่รวมกันระหว่างต้นไม้ตัดสินใจหลาย ๆ ต้น โดยมีการสุ่มคุณลักษณะและข้อมูลตัวอย่างจากชุดข้อมูลเพื่อสร้างแบบจำลอง การคัดเลือกคุณลักษณะได้มีการจัดลำดับความสำคัญของคุณลักษณะโดยใช้ค่า Features Importance เช่นเดียวกับแบบจำลองก่อนหน้า อีกทั้งยังได้ใช้วิธี Boruta ซึ่งเป็นวิธีในการคัดเลือกคุณลักษณะด้วยการสร้างคุณลักษณะเงา (Shadow Features) จากคุณลักษณะดั้งเดิมและเปรียบเทียบความสำคัญโดยการสุ่มสับเปลี่ยน และวิธี BorutaSHAP ซึ่งเป็นวิธีที่เกิดจากการใช้วิธี Boruta ร่วมกับค่า SHAP ในการประเมินกระทบของคุณลักษณะต่อค่าตัวแปรตอบสนอง โดยมีการแบ่งออกเป็น 2 ส่วนคือ การคัดเลือกโดยใช้คุณลักษณะทั้งหมด และการคัดเลือกคุณลักษณะเบื้องต้นโดยใช้ค่า MI Scores ซึ่งเป็นค่าที่ใช้ในการหาค่าความสัมพันธ์ระหว่างคุณลักษณะ โดยทางผู้วิจัยจะเลือกคุณลักษณะที่มีค่า MI Scores 15 อันดับแรกก่อนจะนำมาใช้วิธี Boruta และ BorutaSHAP ในการคัดเลือกคุณลักษณะ ค่า Mi Scores คำนวณได้ตามสมการที่ 2

$$MI(X,Y) = \sum_{x,y} P(x,y) \cdot \log\left(\frac{P(x,y)}{P(x)P(y)}\right) \quad (2)$$

เมื่อ MI(X,Y) คือ ค่า MI Scores ระหว่างตัวแปร X และ Y

P(x,y) คือ คือความน่าจะเป็นที่ X และ Y จะเป็นไปพร้อมกัน

P(x) คือ ความน่าจะเป็นที่ X จะเกิดขึ้น

P(y) คือ ความน่าจะเป็นที่ Y จะเกิดขึ้น

ซึ่งค่า MI Scores เท่ากับ 0 แสดงถึงตัวแปรทั้ง 2 ตัวแปรไม่มีความสัมพันธ์กัน และค่า MI Scores ที่มากขึ้นแสดงถึงความสัมพันธ์ของตัวแปรทั้ง 2 ตัวแปรที่มากขึ้นด้วย

การเพิ่มประสิทธิภาพของแบบจำลอง ผู้วิจัยได้มีการทดลองปรับค่าพารามิเตอร์ (hyperparameter tuning) สำหรับแบบจำลองต่าง ๆ เพื่อที่จะสามารถระบุผลลัพธ์ที่ดีที่สุด ในการประเมินประสิทธิภาพของแบบจำลองได้มีการใช้ค่าคลาดเคลื่อนเฉลี่ยร้อยละแบบสมมาตร (Symmetric Mean Absolute Percentage Error : SMAPE) โดยค่า SMAPE ถูกคำนวณได้โดยสมการที่ 3

$$SMAPE = \frac{100\%}{N} \sum_{t=1}^{N} \frac{|F_t - A_t|}{(|A_t| + |F_t|)/2} \quad (3)$$

เมื่อ SMAPE คือ ค่าคลาดเคลื่อนเฉลี่ยร้อยละแบบสมมาตร

N คือ จำนวนข้อมูลที่ใช้

At คือ ค่าจริงที่เวลา t ใด ๆ

Ft คือ ค่าพยากรณ์ที่เวลา t ใด ๆ

ในการประเมินประสิทธิภาพด้วยการวัดค่า SMAPE นั้น จากโครงสร้างของสมการผลลัพธ์ที่ได้จะให้ค่าออกมาใน

รูปแบบเปอร์เซ็นต์ (%) มีค่าอยู่ระหว่าง 0% ถึง 200% โดยค่า SMAPE ที่ต่ำสุดคือ 0% แสดงถึงความแม่นยำที่สูงที่สุด และค่า SMAPE ที่สูงสุดคือ 200% แสดงถึงความคลาดเคลื่อนที่สูงที่สุด ค่า SMAPE ที่มีค่าอยู่ระหว่าง 0% ถึง 40% จะถูกจัดเป็นช่วงที่มีความแม่นยำสูง ค่าที่อยู่ระหว่าง 40% ถึง 100% จะถูกจัดเป็นช่วงที่มีความแม่นยำที่ยอมรับได้ และค่าที่อยู่ระหว่าง 100% ถึง 200% จะถูกจัดเป็นช่วงที่มีความคลาดเคลื่อนที่สูง หรือไม่เหมาะที่จะนำไปใช้ในการพยากรณ์

4. ผลการวิจัย

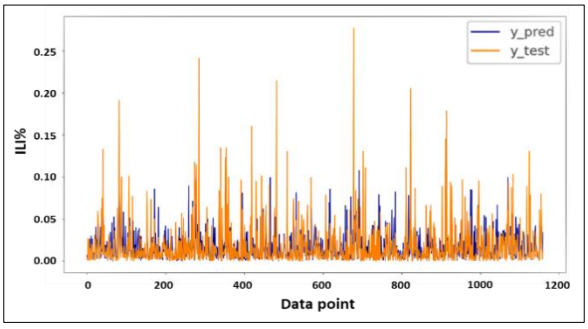
ในการทดลอง ผู้วิจัยได้ทดลองใช้วิธีการคัดเลือกคุณลักษณะสำหรับแบบจำลองการเรียนรู้ของเครื่อง ในตารางที่ 2 แสดงการเปรียบเทียบค่า SMAPE ที่ได้จากการสร้างและทดสอบแบบจำลองด้วยวิธีการคัดเลือกคุณลักษณะต่าง ๆ โดยได้เลือกพารามิเตอร์ที่แสดงประสิทธิภาพดีที่สุดในชุดข้อมูลฝึกสอน และชุดข้อมูลทดสอบ เพียงแบบเดียวของแต่ละแบบจำลองในการนำเสนอในตาราง 2

ตารางที่ 2 การเปรียบเทียบค่า SMAPE ในแบบจำลองด้วยวิธีการคัดเลือกคุณลักษณะต่าง ๆ

Model	Feature Selection	Train set	Test set
		SMAPE	SMAPE
MLR	All Features	86.81%	85.52%
	Stepwise	86.99%	85.16%
Regression Tree	All Features	77.88%	78.18%
	Feature importance	77.88%	78.18%
LightGBM	All Features	50.57%	60.51%
	Feature importance	55.25%	65.21%
XGBoost	All Features	36.40%	60.09%
	Feature importance	37.07%	66.19%
	SHAP	38.01%	61.78%
Random Forest	All Features	39.86%	60.46%
	Feature importance	39.52%	60.96%
	Boruta	39.74%	60.55%
	Boruta + MI Scores	38.13%	59.24%
	BorutaSHAP	39.60%	60.60%
	BorutaSHAP + MI Scores	38.07%	59.11%

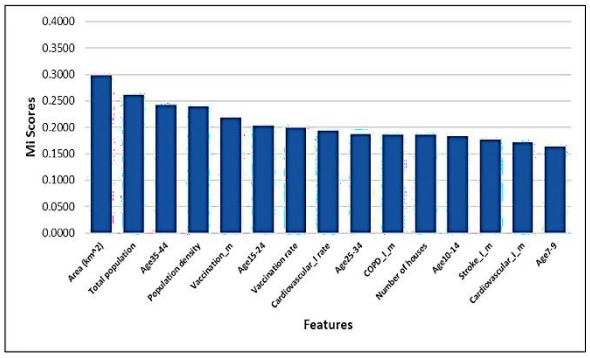
ผลลัพธ์แสดงให้เห็นว่าสมการถดถอยเชิงเส้นพหุคูณ ให้ค่า SMAPE ที่มากสุดซึ่งอาจเป็นเพราะ ปัจจัยที่เกี่ยวข้องนั้นอาจไม่ได้แสดงความสัมพันธ์ในรูปแบบเชิงเส้นตรง ในขณะที่แบบจำลองที่ซับซ้อนขึ้นอย่าง LightGBM , XGBoost และ Random Forest ให้ผลลัพธ์ที่ดีกว่า Regression Tree อย่างเห็นได้ชัด ในแบบจำลอง XGBoost ให้ผลลัพธ์ในชุดข้อมูลฝึกสอนดีที่สุดโดยเฉพาะการใช้คุณลักษณะทุกตัวในการสร้างแบบจำลองพยากรณ์ซึ่งอาจมีผลทำให้อาจเกิดการ Overfitting ได้ แต่ที่น่าสนใจคือแบบจำลอง Random forest ที่มีการใช้ MI scores ในการคัดเลือกคุณลักษณะในเบื้องต้น ให้ผลลัพธ์ในข้อมูลทดสอบดีกว่าแบบจำลองอื่น ๆ รวมถึงการใช้วิธี Boruta และวิธี BorutaSHAP เพียงอย่างเดียว เนื่องจากการใช้ MI scores นั้นมีส่วนช่วยในการคัดเลือกคุณลักษณะที่ไม่เกี่ยวข้องหรือมีความสำคัญน้อยออกจากแบบจำลอง โดยการใช้ MI scores ร่วมกับวิธี BorutaSHAP ให้ผลลัพธ์ดีที่สุดในชุดข้อมูลทดสอบ

การประเมินประสิทธิภาพของ Random forest ที่ใช้วิธี BorutaSHAP ร่วมกับ MI scores ในรูปที่ 3 ได้แสดงการเปรียบเทียบระหว่างค่าจริง (เส้นสีส้ม) และค่าพยากรณ์ (เส้นสีน้ำเงิน) ที่ได้ในชุดข้อมูลทดสอบ อย่างไรก็ตามแบบจำลองมีการประเมินค่าที่ต่ำกว่าความเป็นจริง เนื่องจากค่าของตัวแปร หรือคุณลักษณะที่นำมาใช้ในการสร้างแบบจำลองไม่สามารถตามความแปรปรวนในบริเวณ Peak point ของค่าจริงได้ ซึ่งคุณลักษณะที่มีการจัดความสำคัญจากวิธี BorutaSHAP ร่วมกับ MI scores อย่างอัตราการฉีดวัคซีนนั้น ข้อมูลที่ได้รับมาเป็นข้อมูลรายปี และถูกแจกแจงโดยให้ทุกเดือนมีความสำคัญในระดับที่เท่ากัน ทำให้ผลลัพธ์อาจไม่เป็นที่น่าพอใจในการประเมินความเสี่ยงของการระบาดโรคไข้หวัดใหญ่ในสภาวะฉุกเฉิน แสดงให้เห็นถึงความจำเป็นในการปรับปรุงเพื่อเพิ่มประสิทธิภาพของแบบจำลองในการพยากรณ์โรคไข้หวัดใหญ่ให้แม่นยำยิ่งขึ้น

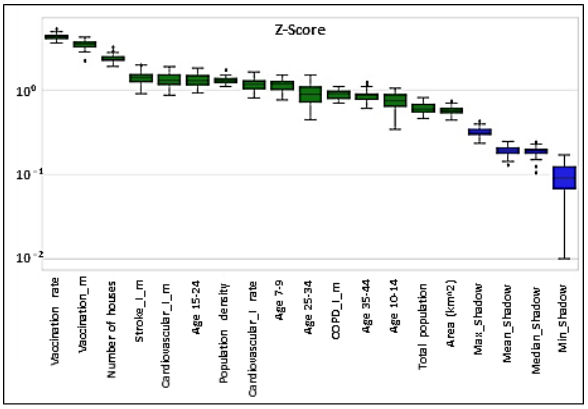


รูปที่ 3 เปรียบเทียบค่า ILI% ระหว่างค่าจริงและค่าพยากรณ์

เพื่อวิเคราะห์คุณลักษณะของแบบจำลองที่ดีที่สุดในชุดข้อมูลทดสอบ คือ แบบจำลอง Random forest ที่ใช้วิธี BorutaSHAP ร่วมกับ MI scores ในรูปที่ 4 แสดงค่า MI scores ของคุณลักษณะที่มีค่ามากสุด 15 อันดับแรก จากนั้นคุณลักษณะที่ถูกคัดเลือก 15 อันดับจากค่า MI scores จะถูกนำมาใช้วิธี BorutaSHAP เพื่อคัดเลือกคุณลักษณะอีกครั้ง ในรูปที่ 5 กราฟแสดงการคัดเลือกคุณลักษณะด้วยค่า Feature importance ของวิธี BorutaSHAP จะเห็นว่าคุณลักษณะที่ถูกคัดเลือกมาทั้ง 15 ตัวถูกยืนยันด้วยวิธี BorutaSHAP ว่ามีสำคัญต่อแบบจำลองโดยการกำหนดสถานะในรูปแบบสี คือ สีเขียวมีความสำคัญต่อแบบจำลอง

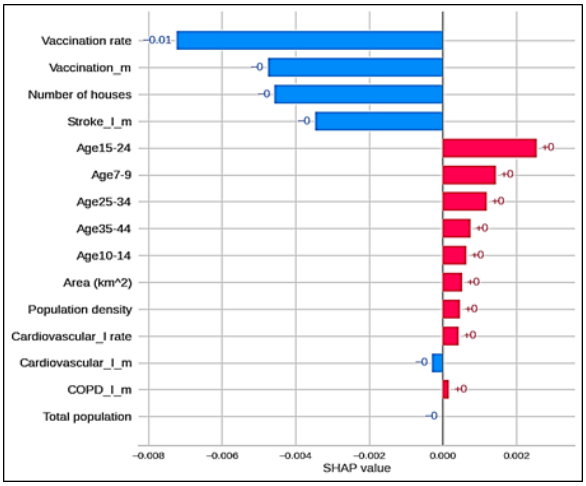


รูปที่ 4 ค่า MI scores ของคุณลักษณะ 15 อันดับแรก



รูปที่ 5 การคัดเลือกคุณลักษณะด้วยค่า Feature Importance ของ BorutaSHAP

ผลลัพธ์อีกอย่างหนึ่งที่ได้จากวิธี BorutaSHAP ร่วมกับ Mi scores คือ ค่า SHAP ที่สามารถแสดงคุณลักษณะที่นำเข้าไปในแบบจำลองว่ามีการเปลี่ยนแปลงการพยากรณ์นั้นอย่างไร ในรูปที่ 6 แสดงค่า SHAP ของแต่ละคุณลักษณะ โดยค่า SHAP ที่มีค่าเป็นค่าลบจะส่งผลเชิงลบต่อค่าพยากรณ์ ซึ่งหมายถึงเมื่อเพิ่มค่าของคุณลักษณะนั้น ๆ จะส่งผลให้ค่าตัวแปรตอบสนองลดลง โดยจะเห็นว่าปัจจัยทางด้านวัคซีน จำนวนบ้าน โรคหลอดเลือดสมองตีบ มีผลต่อค่า ILI% ในเชิงลบ ในขณะเดียวกันค่า SHAP ที่มีค่าเป็นค่าบวกก็จะส่งผลเชิงบวกต่อค่าพยากรณ์ เช่น จำนวนประชากรในช่วงอายุ 15-24 ปี, 7-9 ปี และ 25-34 ปี



รูปที่ 6 ค่า SHAP ของแต่ละคุณลักษณะของวิธี BorutaSHAP ร่วมกับ Mi scores

5. สรุปผลการดำเนินงานและข้อเสนอแนะ

การศึกษานี้ได้ทำการประยุกต์ใช้แบบจำลองต่าง ๆ เพื่อพยากรณ์ แบบจำลองได้ถูกฝึกฝน เปรียบเทียบ และปรับแต่งเพื่อให้ได้ประสิทธิผลที่ดีที่สุด โดยการวิเคราะห์ชุดข้อมูลเกี่ยวกับโรคไข้หวัดใหญ่ การฉีดวัคซีน โรคกลุ่มเสี่ยง สภาพอากาศ และปัจจัยด้านประชากรศาสตร์ ผลลัพธ์แสดงให้เห็นว่า วิธี Random forest ที่มีการคัดเลือกคุณลักษณะในเบื้องต้นโดยใช้ค่า MI score ก่อนการใช้วิธี Boruta หรือ BorutaSHAP เพื่อยืนยันความสำคัญของคุณลักษณะ สามารถพยากรณ์ได้ดีที่สุดด้วยตัวชี้วัด SMAPE เมื่อเทียบกับแบบจำลองอื่น ๆ ที่ใช้คุณลักษณะทั้งหมด หรือ การคัดเลือกคุณลักษณะแบบครั้งเดียว ในการสร้างแบบจำลอง และแสดงคุณลักษณะที่สำคัญในการพยากรณ์ค่า ILI% ได้แก่ อัตราการ

ฉีดวัคซีนต่อจำนวนประชากร จำนวนบ้าน จำนวนผู้ติดเชื้อโรคหลอดเลือดสมองตีบ รวมถึงจำนวนประชากรช่วงอายุ 7-9 ปี และ 15-24 ปี จากข้างต้นที่กล่าวมา การพยากรณ์ค่า ILI% อาจใช้เป็นการประกอบการตัดสินใจในการช่วยเหลือผู้ให้บริการทางการแพทย์ในการจัดสรรวัคซีน หรือ ทรัพยากรที่เกี่ยวข้องเพื่อลดผลกระทบจากการแพร่ระบาดของโรคไข้หวัดใหญ่ ในการพัฒนางานวิจัยในอนาคต ควรพิจารณาการเพิ่มประสิทธิภาพของแบบจำลองด้วยการหาคุณลักษณะหรือปัจจัยที่เกี่ยวข้องเพิ่มเติม เช่น อัตราการฉีดวัคซีน และอัตราการติดเชื้อตามช่วงอายุต่าง ๆ การเพิ่มความละเอียดของข้อมูลที่มากขึ้น และการเก็บข้อมูลที่มีการเจาะจงในระดับชุมชน การใช้แบบจำลองเชิงลึกที่มีความซับซ้อนมากยิ่งขึ้นอาจช่วยในการปรับปรุงประสิทธิภาพของการพยากรณ์ นอกจากนี้ การพิจารณาข้อมูลปัจจัยทางสังคม อย่างเช่น พฤติกรรมของคนในชุมชน กิจกรรมทางสังคม อาจเป็นปัจจัยที่ควรนำมาพิจารณาร่วมด้วยในการนำไปใช้วิเคราะห์เพื่อบรรเทาผลกระทบของโรคระบาด

6. เอกสารอ้างอิง

[1] ณัฐพล จำปาสาร, รายงานประจำปี 2565: Influenza, กรุงเทพมหานคร: กรมควบคุมโรค สำนักงานควบคุมโรคที่ 9 กระทรวงสาธารณสุข, 2565.

[2] H. N. Zhai, “The Analysis of Relationship Between Influenza and Atmospheric Ambient and the Establishment of Influenza-Like-Illness Rates Forecasting Model,” China University of Geosciences Wuhan, China, 2009.

[3] D. Sanguanrungsirikul, H. Chiewanantavanich and M. Sangkasem, "A Comparative study to determine optimal models for forecasting the number of patients having epidemiological-surveillance diseases in Bangkok," *KMUTT Research and Development Journal.*, vol. 38, no. 1, pp. 35-55, 2015.

[4] M. J. Paul, M. Dredze and D. Broniatowski, “Twitter Improves Influenza Forecasting,” *PLOS Currents.*, vol. 1, no. 6, 2014.

[5] S. Kandula and J. Shaman, “Near-term forecasts of influenza-like illness,” *Epidemics.,* vol. 27, pp. 41–51, 2019.

[6] S. Maairkien, D. Areechokchai, S. Saita and T. Silawan, “Time series analysis and forecast of Influenza cases for different age groups in Phitsanulok Province, Northern Thailand,” *Current Applied Science and Technology.*, vol. 20, no. 2, pp. 310–320, 2020.

[7] E. L. Ray and N. G. Reich, “Prediction of infectious disease epidemics via weighted density ensembles,” *PLOS Computational Biology.*, vol. 14, no. 2, pp. 1-23, 2018.

[8] Y. Lu, S. Wang, J. Wang, G. Zhou, Q. Zhang, X. Zhou, B. Niu, Q. Chen and K. C. Chou, “An epidemic Avian Influenza prediction model based on Google Trends,” *Letters in Organic Chemistry.,* vol. 16, no. 4, pp. 303-310, 2019.

[9] X. Zhou, Y. Zhang, C. Shen, A. Liu, Y. Wang, Q. Yu, F. Guo, A. C. A. Clements, C. Smith, J. Edwards, B. Huang and R. J. S. Magalhães, “Knowledge, attitudes, and practices associated with Avian Influenza along the live chicken market chains in Eastern China: a cross-sectional survey in Shanghai, Anhui, and Jiangsu,” *Transboundary and Emerging Diseases.,* vol. 66, no. 4, pp. 1529-1538, 2019.

[10] H. Xue, Y. Bai, H. Hu and H. Liang, “Influenza activity surveillance based on multiple regression model and artificial neural network,” *IEEE Access.*, vol. 6, pp. 563–575, 2018.

[11] H. Xue, L. Zhang, H. Liang, L. Kuang, H. Han, X. Yang and L. Guo, “Influenza trend prediction method combining Baidu index and support vector regression based on an improved particle swarm optimization algorithm,” *AIMS Mathematics.*, vol. 8, no. 11, pp. 25528-25549, 2023.

[12] J. Zhang and K. Nawata, “Multi-step prediction for influenza outbreak by an adjusted long short-term

memory," *Epidemiology and Infection.,* vol. 146, no. 7, pp. 809–816, 2018.

[13] A. Kara, "Multi-step influenza outbreak forecasting using deep LSTM network and genetic algorithm," *Expert Systems with Applications.*, vol. 180, no. 3, 2021.

[14] H. Stoppiglia, G. Dreyfus, R. Dubois and Y. Oussar, "Ranking a random feature for variable and feature selection," *Journal of Machine Learning Research.*, vol. 3, pp. 1399–1414, 2003.

[15] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *31st Conference on Neural Information Processing Systems*, United States, 2017.