วารสารวิทยาศาสตร์บูรพา ปีที่ 28 (ฉบับที่ 1) มกราคม – เมษายน พ.ศ. 2566
BURAPHA SCIENCE JOURNAL Volume 28 (No.1) January – April 2023 บทความวิจัย

การประยุกต์ใช้วิศวกรรมคุณลักษณะและตัวแบบเชิงเส้นนัยทั่วไป สำหรับพยากรณ์จำนวนผู้ติดเชื้อใหม่ไวรัสโคโรนา 2019

An Application of Feature Engineering and Generalized Linear Model for Forecasting the Number of COVID-19 New Cases

ณฐกร นวรตน[1], พันงาม วงศ์คำจันทร์[2], สุขเกษม วัชรมัยสกุล[2] และ เจษฎา ตัณฑนุช[1*]

Natakon Nawaratana[1], Punngam Wongcumjan[2], Sukasem Watcharamaisakool[2] and Jessada Tanthanuch[1*]

[1] *สาขาวิชาคณิตศาสตร์ สำนักวิชาวิทยาศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี*

[2] *สาขาวิชานวัตกรรมวิศวชีวการแพทย์ สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี*

[1]*School of Mathematics, Institute of Science, Suranaree University of Technology*

[2]*School of Biomedical Innovation Engineering, Institute of Engineering, Suranaree University of Technology*

Received : 5 July 2022

Revised : 30 August 2022

Accepted : 4 October 2022

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อสร้างตัวแบบเชิงเส้นนัยทั่วไปเพื่อพยากรณ์จำนวนผู้ติดเชื้อไวรัสโคโรนา 2019 ที่จะเกิดขึ้นใหม่ การดำเนินการวิจัยใช้ข้อมูลสาธารณะ COVID-19 Dataset ของ DEVAKUMAR ปรับปรุงเมื่อวันที่ 30 มกราคม 2563 จากเว็บไซต์ www.kaggle.com ซึ่งข้อมูลดังกล่าวเป็นข้อมูลเกี่ยวกับผู้เชื้อไวรัสโคโรนา 2019 จาก 187 ประเทศประกอบด้วยตัวแปรตอบสนอง 1 ตัวแปร และตัวแปรอธิบาย 12 ตัวแปร ในการดำเนินการวิจัยได้ประยุกต์ใช้วิธีวิศวกรรมคุณลักษณะอัตโนมัติทำให้ลดการใช้ตัวแปรอธิบายเหลือเพียง 6 ตัวแปร แต่มีคุณลักษณะที่มีนัยสำคัญในการสร้างตัวแบบ 7 คุณลักษณะ ได้แก่ จำนวนคนที่เสียชีวิตใหม่ จำนวนคนที่ติดเชื้อในรอบสัปดาห์ จำนวนคนที่หายจากการติดเชื้อสะสม จำนวนคนที่หายจากการติดเชื้อใหม่ จำนวนคนที่ติดเชื้อสะสม จำนวนคนที่อยู่ระหว่างการรักษา และผลคูณระหว่างจำนวนคนที่หายจากการติดเชื้อใหม่กับจำนวนคนที่อยู่ระหว่างการรักษา จากนั้นนำข้อมูลคุณลักษณะดังกล่าวไปดำเนินการสร้างตัวแบบด้วยวิธีตัวแบบเชิงเส้นนัยทั่วไป โดยตั้งสมมติฐานว่าข้อมูลมีรูปแบบการแจกแจงทางสถิติ 3 รูปแบบ ได้แก่ การแจกแจงปรกติ การแจกแจงทวินามลบ และการแจกแจงปัวซง ขั้นตอนถัดมานำตัวแบบที่ได้ไปปรับปรุงเพื่อเพิ่มประสิทธิภาพโดยการใช้กระบวนการการคัดเลือกตัวแปรแบบลำดับขั้น ผลการศึกษาพบว่าตัวแบบเชิงเส้นนัยทั่วไปที่ใช้การแจกแจงปัวซงเป็นตัวแบบที่มีประสิทธิภาพดีที่สุด โดยตัวแบบใช้ทั้ง 7 คุณลักษณะในการสร้างและมีค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 365.0387 และค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ยเท่ากับ 803.0267 ขณะที่ตัวแบบเชิงเส้นนัยทั่วไปที่ใช้การแจกแจงปรกติมีประสิทธิFigureต่ำกว่าเล็กน้อย โดยมีค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 365.4591 และค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ยเท่ากับ 803.0286 แต่ใช้เพียง 4 คุณลักษณะเท่านั้นในการสร้างตัวแบบ ซึ่งได้แก่ จำนวนคนที่เสียชีวิตใหม่ จำนวนคนที่ติดเชื้อในรอบสัปดาห์ จำนวนคนที่หายจากการติดเชื้อใหม่ และจำนวนคนที่อยู่ระหว่างการรักษา ผลการดำเนินการที่ได้ช่วยให้ได้กระบวนทัศน์ของการประยุกต์ใช้วิธีวิศวกรรมคุณลักษณะมาลดความซับซ้อนในการสร้างตัวแบบเชิงเส้นนัยทั่วไปสำหรับพยากรณ์

คำสำคัญ : ไวรัสโคโรนา 2019 ; ตัวแบบเชิงเส้นนัยทั่วไป ; วิศวกรรมคุณลักษณะ

Abstract

The purpose of this research is to construct a generalized linear model (GLM) for forecasting the number of new COVID-19 cases. The data used in this research is the open-source COVID-19 dataset from DEVAKUMAR updated on January 30, 2020, from www.kaggle.com. The dataset contains information of infected COVID-19 patients data collected from 187 countries and is composed of 1 responsive variable and 12 explanatory variables. Through feature engineering, it was found that there were 6 significant explanatory variables only. These variables provided 7 significant features, which were the number of new deaths, number of new cases in a week, number of recovered cases, number of newly recovered cases, number of confirmed cases, number of active cases, and the product of the number of new recovered cases with the number of active cases. The 7 features were used to create the GLM under the assumption that the data might be classified following one of these three statistical distributions, normal distribution, negative binomial distribution, and Poisson distribution. After that, the models were modified for improving their performance by using the stepwise selection technique. The study showed that the GLM by Poisson distribution provided the best performance. By using all 7 features, the model by Poisson distribution has RMSE = 365.0387 and MAE = 803.0267. However, the GLM by normal distribution provided a marginally lower performance, RMSE = 365.4591 and MAE = 803.0286, by using 4 features only. The 4 features used for modeling were the number of new deaths, number of new cases in a week, number of newly recovered cases, and number of active cases. The result of this implementation allows for a paradigm of applying feature engineering methods to simplify the creation of generalized linear models for forecasting.

Keywords : Covid-19 ; generalized linear model ; feature engineering

***Corresponding author. E-mail :** jessada@g.sut.ac.th

บทนำ

องค์การอนามัยโลก (World Health Organization - WHO) ได้ประกาศให้โรคที่เกิดจากไวรัสโคโรนาหรือโรคโควิด-19 (COVID-19) เป็นภาวะการระบาดใหญ่ (Coronavirus Pandemic) เมื่อวันที่ 11 มีนาคม 2563 (Emerging Infectious Disease Work of Communicable Disease Academic Development Group, 2021)เนื่องด้วยเชื้อไวรัสโคโรนาได้พร่กระจายไปทั่วทุกมุมโลกอย่างรวดเร็ว การแพร่ระบาดในครั้งนี้ได้สร้างความสูญเสียครั้งยิ่งใหญ่ให้กับมวลมนุษยชาติ ประชากรล้มป่วยและเสียชีวิตเป็นจำนวนมาก การแพร่ระบาดยังได้ส่งผลกระทบเป็นวงกว้างในทุกมิติรวมไปถึงทางด้านการศึกษา เศรษฐกิจและสังคม เกิดการหยุดเรียนและมีการเรียนผ่านระบบออนไลน์ ห่วงโซ่ธุรกิจถูกทำลายเป็นวงกว้าง การลงทุนถดถอย การค้าขายฝืดเคือง การท่องเที่ยวตกต่ำ ภาคอุตสาหกรรมไม่สามารถดำเนินการได้ ทำให้แรงงานตกงาน ผู้คนกลับสู่ภาวะยากจนเรื้อรัง (Amattayakul, 2020) และผลกระทบดังกล่าวอาจจะลุกลามทำให้ระบบกลไกการพัฒนาประเทศของประเทศไทยเกิดการเสียสมดุล ดังนั้นสำนักงานสภาพัฒนาการเศรษฐกิจและสังคมแห่งชาติ(สศช.) จึงให้ประกาศแผนแม่บทเฉพาะกิจภายใต้ยุทธศาสตร์ชาติ อันเป็นผลมาจากสถานการณ์โควิด-19 พ.ศ. 2564-2565 ในราชกิจจานุเบกษา ภายใต้แนวคิด “ล้มแล้วลุกไว” เพื่อให้มาตรการการพัฒนาประเทศได้กลับมาขับเคลื่อนได้อย่างต่อเนื่องอีก (Strategy and Organization Development Group, 2022) และคณะกรรมการโรคติดต่อแห่งชาติได้มีมติเห็นชอบกับแนวทางการปรับให้โรคโควิด-19 ออกจากโรคระบาดเข้าสู่โรคประจำถิ่น เพื่อเป็นการบรรเทาผลกระทบที่เกิดขึ้นจากการระบาดนี้ โดยมีเป้าหมายจะเริ่มตั้งแต่วันที่ 1 กรกฎาคม 2565 เป็นต้นไป (Department of Mental Health, 2022)

การออกแบบนโยบายควบคุมโรคหรือแผนการจัดเตรียมทรัพยากรในการรองรับการระบาด ซึ่งได้แก่ สถานที่เวชภัณฑ์ บุคลากรทางการแพทย์ และข้อกำหนดกฎหมายต่าง ๆ จำเป็นต้องประเมินจากรูปแบบการแพร่ระบาดของโรคโควิด-19 ซึ่งเป็นโรคทางเดินหายใจที่เกิดจากของเชื้อไวรัส SARS-CoV-2 ที่มีลักษณะแพร่เชื้อจากคนสู่คน ผ่านทางฝอยละอองของสารคัดหลั่งจากทางจมูกหรือปาก เมื่อได้รับเชื้อจะมีระยะฟักตัวซึ่งเป็นช่วงที่ยังไม่แสดงอาการ ดังนั้นผู้ติดเชื้อมักแพร่เชื้อจากการไอ จาม หรือสัมผัส ทำให้เชื้อโรคลอยกระจายทางอากาศและฝังตัวตามสิ่งของรอบข้าง ซึ่งเป็นผลให้ผู้คนที่อยู่บริเวณใกล้เคียงโดยรอบอาจได้รับเชื้อโรคเข้าสู่ร่างกายโดยไม่รู้ตัว (Leelarutsamee, n.d.) จึงเป็นสาเหตุที่โรคโควิด-19 มีอัตราการระบาดที่เพิ่มขึ้นรวดเร็วเป็นอย่างมาก โดย Kasilingam *et al.* (2021) ได้เลือกใช้การสร้างแบบจำลองการเติบโตเลขชี้กำลัง (exponential growth modelling) ที่สอดคล้องกับรูปแบบอัตราการเพิ่มขึ้นของจำนวนผู้ติดเชื้อแบบก้าวกระโดด ร่วมกับเทคนิคการเรียนรู้เครื่อง (machine learning) ในการทำนายการแพร่กระจายของโรคระบายโควิด-19 จากข้อมูลผู้ติดเชื้อใน 42 ประเทศ

ตัวแบบเชิงเส้นนัยทั่วไป (General Linear Model, GLM) เป็นเครื่องมือทางทางคณิตศาสตร์และสถิติที่มักจะได้รับความสนใจเป็นลำดับแรก ๆ ถ้าต้องทำงานร่วมกับข้อมูลที่มีการแจกแจงทางสถิติในวงศ์เลขชี้กำลัง (exponential family) ตัวแบบดังกล่าว เป็นการสร้างแบบจำลองทางคณิตศาสตร์เพื่อใช้ในการหาแนวโน้มความสัมพันธ์ระหว่างตัวแปรต้นกับตัวแปรตาม ที่ง่ายต่อการทำความเข้าใจและได้รับความนิยมอย่างกว้างขวางในการวิเคราะห์ข้อมูลที่เกี่ยวข้องกับผู้ติดเชื้อไวรัสโคโรนา 2019 เนื่องจากตัวแบบสามารถปรับโครงสร้างได้และมีความยืดหยุ่น สามารถนำไปใช้กับการถดถอย (regression) ในรูปแบบต่าง ๆ ได้ และมีโปรแกรมสำเร็จรูปจำนวนมากที่ช่วยในการสร้างตัวแบบ เช่น Xie & Farrell (2020) ได้วิเคราะห์ชุดข้อมูลของ

ผู้ที่ได้รับการยืนยันว่าติดเชื้อ ผ่านโปรแกรม RStudio ซึ่งเป็นซอฟต์แวร์ได้ช่วยสร้างและปรับปรุงแบบจำลองการถดถอยที่ใช้การแจกแจงทางสถิติแบบปัวซง (Poisson distribution) ในปีเดียวกัน Benlagha (2020) ได้เปรียบเทียบแบบจำลองทางคณิตศาสตร์โดยวิธีการถดถอยบนสมมติฐานว่าตัวแปรสุ่มมีการแจกแจงความน่าจะเป็นแบบ ปัวซงและแบบทวินามลบ (negative binomial distribution) เพื่อพยากรณ์จำนวนของผู้ป่วยรายใหม่ของโรคโควิด-19 ของ 8 ประเทศที่ได้รับผลกระทบสูงสุดจากการระบาด ต่อมาในด้านระบาดวิทยา Vytla *et al*. (2021) ได้มีการสร้างแบบจำลองการแพร่ระบาดของเชื้อที่มีความไม่แน่นอน ซึ่งพบว่ารูปแบบการแพร่ระบาดมีความคล้ายกับการสุ่มโดยธรรมชาติในลักษณะการแจกแจงแบบเกาส์ (Gaussian distribution) หรือเรียกอีกอย่างหนึ่งว่าการแจกแจงแบบปรกติ (normal distribution)

ในปีปัจจุบันมีการพัฒนาตัวแบบโดยการเพิ่มขั้นตอนจัดการกับคุณลักษณะ (feature) ที่มีอิทธิพลกับการสร้างตัวแบบด้วยเทคนิควิศวกรรมคุณลักษณะ (feature engineering) ซึ่งเป็นกระบวนการที่ใช้โดเมนของข้อมูลเพื่อที่จะเรียนรู้ในการหาสหสัมพันธ์ (correlation) ระหว่างตัวแปรต่าง ๆ ในตัวแบบ การประยุกต์ใช้การจัดการคุณลักษณะมีประโยชน์ที่จะสามารถช่วยลดระยะเวลาในการทำงาน และเพิ่มประสิทธิภาพการตัดสินใจในการสร้างตัวแบบทำนายจำนวนผู้เสียชีวิตของโรคโควิด-19 รายใหม่ที่เพิ่มขึ้น ได้อย่างก้าวกระโดด (Kelter *et al*, 2021)

จากข้อมูลข้างต้นงานวิจัยนี้ต้องการประยุกต์ใช้เทคนิควิศวกรรมคุณลักษณะเพื่อจัดการกับคุณลักษณะหรือตัวแปรแล้วดำเนินการการสร้างตัวแบบทางคณิตศาสตร์ในรูปแบบตัวแบบเชิงเส้นนัยทั่วไป ทั้งนี้พิจารณาการสร้างตัวแบบจากการแจกแจงทางสถิติแบบปัวซง แบบทวินามลบ และแบบปรกติ จากนั้นใช้การคัดเลือกคุณลักษณะแบบลำดับขั้น (stepwise selection) เพื่อหาเซตย่อยของคุณลักษณะที่เหมาะสมที่ทำให้ตัวแบบมีประสิทธิภาพสูงสุด แล้วนำไปทดสอบกับข้อมูลจำนวนผู้ติดเชื้อรายใหม่ของโรคโควิด-19 เพื่อเปรียบเทียบประสิทธิภาพของตัวแบบที่ถูกสร้างขึ้น ซึ่งการได้ตัวแบบที่มีประสิทธิภาพจะช่วยให้การวางแผนการจัดเตรียมทรัพยากรในการรองรับการระบาด การกำหนดนโยบายและออกกฎหมายต่าง ๆ ที่เกี่ยวข้องทำได้อย่างเหมาะสมกับสถานการณ์ และช่วยให้ก้าวผ่านวิกฤตการระบาดของโรคได้อย่างรวดเร็ว

<u>ทฤษฎีและองค์ความรู้ที่เกี่ยวข้อง</u>

วิศวกรรมคุณลักษณะ (Feature engineering) คือ กระบวนการที่ใช้โดเมนของข้อมูลสร้างคุณลักษณะเพื่อให้การเรียนรู้ของเครื่องทำงานได้ตามขั้นตอน การจัดการคุณลักษณะเป็นพื้นฐานของการประยุกต์ ใช้การเรียนรู้ของเครื่อง

กระบวนการทำงานของการจัดการคุณลักษณะ (javaTpoint, 2021)

1) ศึกษาและทดสอบรูปแบบของคุณลักษณะ
2) ตัดสินใจเลือกคุณลักษณะที่จะสร้าง
3) สร้างคุณสมบัติ
4) ตรวจสอบและทดสอบคุณสมบัติที่สร้างขึ้น
5) ปรับปรุงคุณสมบัติ (หากจำเป็น)
6) กลับไปที่ขั้นที่ 1) เพื่อสร้างคุณลักษณะใหม่หรือเพิ่มเติมจนกว่าจะได้ค่าคลาดเคลื่อนที่ยอมรับได้

วิศวกรรมคุณลักษณะเป็นขั้นตอนวิธีที่ง่ายต่อการทำงาน อีกทั้งช่วยให้การเลือกตัวแปรเพื่อการทำนาย ที่มีนัยสำคัญสำหรับการสร้างตัวแบบเป็นไปอย่างมีประสิทธิภาพ ทำให้การสร้างตัวแบบที่ต้องการเป็นไปได้ง่ายขึ้น

วิศวกรรมคุณลักษณะแบบอัตโนมัติ (Automatic Feature Engineering) เนื่องด้วยการจัดการคุณลักษณะมีผลต่อการสร้างตัวแบบ โดยที่บางครั้งไม่จำเป็นต้องใช้คุณลักษณะที่ได้มาทั้งหมด เราสามารถพิจารณาใช้เฉพาะคุณลักษณะที่มีนัยสำคัญเท่านั้นมาดำเนินการสร้างตัวแบบก็ได้ ทำให้ลดความซับซ้อนในการสร้างตัวแบบลง ทั้งนี้เทคนิควิศวกรรมคุณลักษณะแบบอัตโนมัติไม่เพียงแต่คัดเลือกคุณลักษณะที่มีนัยสำคัญต่อการสร้างตัวแบบเท่านั้น แต่ยังสามารถสร้างคุณลักษณะเพิ่มเติมโดยการนำคุณลักษณะที่มี มาดำเนินการทางคณิตศาสตร์ เช่น ดำเนินการทางฟังก์ชันตรีโกณมิติ ลอการิทึม หรือนำคุณลักษณะที่มีอยู่มาดำเนินการทางพีชคณิตศาสตร์บวก ลบ คูณ หาร หรือ ยกกำลัง ซึ่งเรียกการดำเนินการส่วนนี้ว่ากระบวนการสร้างคุณลักษณะใหม่ (wrapper method) แล้วพิจารณาว่ามีนัยสำคัญต่อการสร้างตัวแบบหรือไม่ เพื่อใช้เป็นส่วนหนึ่งในการสร้างตัวแบบ (Reis, 2019)

การทดสอบประสิทธิภาพแบบไขว้ (Cross Validation) คือวิธีการสุ่มแบ่งชุดข้อมูลออกเป็นส่วน ๆ ที่เท่ากันและสุ่มนำเอากลุ่มตัวอย่างที่ถูกแบ่งบางส่วนมาตรวจสอบค่าความคลาดเคลื่อนกับข้อมูลชุดอื่นวนทดสอบไปจนครบตามจำนวนที่ถูกแบ่งไว้ และผลเฉลี่ยของค่าความคลาดเคลื่อนทั้งหมดทุกรอบจะเป็นเครื่องมือที่ช่วยเพิ่มความน่าเชื่อถือในการสร้างตัวแบบยิ่งขึ้น เช่น 10-fold cross-validation คือ การแบ่งข้อมูลออกเป็น 10 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูล เท่ากัน หลังจากนั้นข้อมูลหนึ่งส่วนจะใช้เป็นชุดทดสอบประสิทธิภาพของแบบจำลอง เรียนรู้วนไปจนกว่าข้อมูลจะถูกใช้ครบทุกส่วนตามจำนวนที่กำหนด แล้วนำมาหาค่าเฉลี่ย ทำให้ได้ผลลัพธ์ที่มีความน่าเชื่อถือ (Patcharawongsakda, 2014)

เกณฑ์สารสนเทศของอะกะอิเกะ หรือ เอไอซี (Akaike Information Criterion - AIC) เป็นเครื่องมือในการประเมินคุณภาพของตัวแบบเชิงสถิติโดยประเมินจากค่าคลาดเคลื่อนในการทำนายนอกช่วงการประมาณค่า คำนวณได้จาก

$$\text{AIC} = -2\ln\hat{L} + 2k$$

โดยที่ $\hat{L}$ คือ ค่าสูงสุดของฟังก์ชันภาวะน่าจะเป็น (likelihood function) ของตัวแบบ และ k คือจำนวนตัวแปรหรือพารามิเตอร์ในตัวแบบ (Office of the Royal Thai Embassy, 2018) โดยปกติแล้วตัวแบบที่มีความเหมาะสมในการอธิบายตัวแปรตอบสนองนั้นจะต้องการค่าเอไอซีที่ต่ำที่สุด

การคัดเลือกตัวแปรแบบลำดับขั้น (Stepwise Selection) ในการทำการคัดเลือกตัวแปรเพื่อปรับปรุงประสิทธิภาพของตัวแบบมีแนวทางปฏิบัติในการคัดเลือกตัวแปรแบบลำดับขั้น 3 แนวทางได้แก่

1) Forward stepwise วิธีนี้จะเริ่มจากการนำตัวแปรที่มีแสดงความสัมพันธ์มากที่สุดเพิ่มเข้าสู่สมการทีละ 1 ตัว โดยตัวแปรที่ทำให้ค่า $-2\ln\hat{L}$ ของตัวแบบลดลง แสดงว่าตัวแปรนั้นควรจะคงอยู่ในตัวแบบ
2) Backward stepwise เป็นวิธีที่ทำตรงกันข้ามกับวิธี Forward stepwise คือพิจารณาว่าจะนำตัวแปรตัวใดไม่มีผลต่อการพยากรณ์หรือความสัมพันธ์น้อยที่สุดออกจากตัวแบบทีละ 1 ตัว และ

3) Both stepwise แนวทางนี้ในทุกครั้งจะทั้งการคัดเลือกเพิ่มตัวแปรแบบ Forward stepwise และตัดตัวแปรแบบ Backward stepwise สลับกันไปในการทำตัวแบบ (Nawaratana, 2019)

ตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model - GLM) เป็นตัวแบบที่ขยายจากตัวแบบเชิงเส้นไปสู่ตัวแบบที่การแจกแจงของตัวแปรตอบสนองในตัวแบบอยู่ในวงศ์เลขชี้กำลัง

พิจารณาตัวแบบการถอยเชิงเส้น (linear regression model) โดยใช้สัญลักษณ์ตัวแปรอธิบาย (explanatory variable: X) กับตัวแปรตอบสนอง (response variable: Y) บนสมมติฐาน

$$Y = X\beta + \varepsilon$$

เมื่อ n คือจำนวนตัวแปรอธิบาย, k คือจำนวนตัวแปรตอบสนอง, $Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ เป็นเวกเตอร์แนวตั้งของตัวแปรตอบสนอง, $X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix}$ เป็นเมทริกซ์ของข้อมูลที่จะใช้สร้างตัวแบบเชิงเส้นนัยทั่วไป, $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{k+1} \end{bmatrix}$ เป็นเวกเตอร์แนวตั้งของสัมประสิทธิ์การถดถอย (regression coefficient) $\beta_i, i = 1, \dots, k+1$ และ $\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$ เป็นเวกเตอร์แนวตั้งค่าคลาดเคลื่อนของตัวแบบเชิงเส้นนัยทั่วไป โดย $\varepsilon_i, i = 1, \dots, n$ เป็นตัวแปรสุ่มที่มีการแจกแจงปกติ โดยมีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนคือ δ^2

จุดประสงค์ของการสร้างตัวแบบเชิงเส้นวางนัยทั่วไป คือ การหาสัมประสิทธิ์การถดถอยที่ทำให้ความสัมพันธ์ระหว่างตัวแปรอธิบาย X กับค่าคาดหมาย (expectation) ของตัวแปรตอบสนอง Y มีความคลาดเคลื่อนน้อยที่สุด โดยตัวแบบเชิงเส้นนัยทั่วไปประกอบไปด้วย 3 ส่วนได้แก่ 1) ส่วนประกอบเชิงสุ่ม (random component) ซึ่งในที่นี้คือ Y 2) ส่วนประกอบเชิงระบบ (systematic component) ซึ่งเป็นตัวผลรวมเชิงเส้น ซึ่งในที่นี้คือ $X\beta$ และ 3) ฟังก์ชันเชื่อมโยง (link function) ซึ่งเป็นฟังก์ชันของค่าคาดหมายของตัวแปรตอบสนอง ทั้งนี้ตัวแบบเชิงเส้นนัยทั่วไปคือ

$$g\big(E(Y)\big) = X\beta,$$

เมื่อ $E(Y)$ ค่าคาดหมายของตัวแปรตอบสนอง Y และ $g(\cdot)$ คือฟังก์ชันเชื่อมโยง เช่น ฟังก์ชันเชื่อมโยงเอกลักษณ์ $g(\mu) = \mu$ ฟังก์ชันเชื่อมโยงส่วนกลับ $g(\mu) = \frac{1}{\mu}$ และ ฟังก์ชันเชื่อมโยงลอการิทึม $g(\mu) = \ln \mu$ (Office of the Royal Thai Embassy, 2018)

การประเมินประสิทธิภาพ (Performance Evaluation) ในงานวิจัยนี้จะประเมินประสิทธิภาพของตัวแบบด้วยค่าคลาดเคลื่อนต่อไปนี้

ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Error: MAE) $\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\text{Error}|$

และ ค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Square Error: RMSE) $\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{Error})^2}$ ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย MAE จะแสดงให้เห็นถึงค่าเฉลี่ยของขนาด (magnitude) ของความคลาดเคลื่อน แต่สำหรับค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย RMSE เป็นการหาค่ารากที่สองของผลรวมกำลังสองของขนาดความคลาดเคลื่อนซึ่งแสดงถึงค่าเฉลี่ยของความคลาดเคลื่อนเช่นกัน แต่หากค่าความคลาดเคลื่อนของข้อมูลมีการกระจายตัวผิดปรกติ ค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย RMSE จะสามารถแสดงให้เห็นถึงความคลาดเคลื่อนของตัวแบบได้ชัดเจนกว่าค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย MAE ทั้งนี้หากค่าคลาดเคลื่อนทั้งสองของตัวแบบน้อยกว่า จะถือว่าตัวแบบมีประสิทธิภาพในการพยากรณ์ที่ดีกว่า ค่าคลาดเคลื่อนทั้งสองสามารถคำนวณได้ง่ายและช่วยให้สะดวกต่อการแปลผลในการประเมินประสิทธิภาพของตัวแบบ

โดยภาพรวมการดำเนินการวิจัยนี้ต้องการใช้วิศวกรรมคุณลักษณะเพื่อวิเคราะห์คุณลักษณะที่เหมาะสมสำหรับการสร้างตัวแบบเชิงเส้นวางนัยทั่วไป โดยใช้การประเมินความเหมาะสมด้วยเอไอซีและการทดสอบประสิทธิภาพแบบไขว้ จากนั้นใช้การการคัดเลือกตัวแปรแบบลำดับขั้นทำการคัดเลือกคุณลักษณะในการสร้างตัวแบบ และจะประเมินประสิทธิภาพของตัวแบบที่ได้ด้วยค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย MAE และค่าคลาดเคลื่อนกำลังสองเฉลี่ย RMSE

วิธีดำเนินการวิจัย

ในการดำเนินการวิจัยนี้ได้ใช้ข้อมูลจากฐานข้อมูลสาธารณะ https://www.kaggle.com/imdevskp/corona-virus-report ซึ่งเป็นฐานข้อมูล COVID-19 Dataset ของ DEVAKUMAR ปรับปรุงเมื่อวันที่ 30 มกราคม 2563 โดยเป็นข้อมูลผู้ติดเชื้อไวรัสโคโรนา 2019 จำนวน 187 ประเทศ มีผู้สนใจนำไปใช้ทำงานวิจัย 1721 คน โดยรวบรวมข้อมูลจาก https://github.com/CSSEGISandData/COVID-19 และ https://www.worldometers.info/ ข้อมูลที่ได้เป็นแฟ้มข้อมูลประเภท CSV ประกอบไปด้วย 13 ตัวแปร ดังแสดงในTable 1 และเนื่องด้วยโปรแกรม RStudio เป็นโปรแกรมที่ถูกยอมรับในด้านการทำงานทางด้านสถิติอย่างกว้างขวางและมีศักยภาพในการสร้างตัวแบบเชิงเส้นวางนัยทั่วไปสูง งานวิจัยนี้จะใช้โปรแกรม RStudio version 3.6.1 ภายใต้ ระบบปฏิบัติการ Microsoft Windows 10 Pro รุ่น 21H2 เป็นโปรแกรมหลักในการสร้างตัวแบบและดำเนินการคำนวณบนเครื่องคอมพิวเตอร์ ที่ใช้ CPU Intel รุ่น I5-6200U แต่เนื่องด้วยโปรแกรม RStudio ไม่มีคำสั่งทางด้านวิศวกรรมคุณลักษณะแบบอัตโนมัติ ดังนั้นการดำเนินการเกี่ยวกับการจัดการคุณลักษณะจะดำเนินการโดยโปรแกรม RapidMiner Studio 9.2.0 (Education License)

ขั้นตอนที่ 1 การวิเคราะห์หาคุณลักษณะเพื่อการคัดกรองตัวแปรด้วยโปรแกรม RapidMiner Studio

สำหรับในงานวิจัยครั้งนี้จะใช้การดำเนินการวิศวกรรมคุณลักษณะแบบอัตโนมัติ (automatic feature engineering) ซึ่งเป็นการดำเนินการเลือกใช้กระบวนการทางคณิตศาสตร์เพื่อคัดกรองคุณลักษณะในการสร้างตัวแบบ กระบวนการดังกล่าวจะพยายามศึกษาหาความเชื่อมโยงของคุณลักษณะขึ้นมาหลายรูปแบบที่มีนัยสำคัญ แล้วสกัดคุณลักษณะที่ส่งผลต่อการสร้างตัวแบบของข้อมูลนั้น ถ้าหากตัวแปรไหนไม่มีประโยชน์หรือมีความสำคัญน้อยมากต่อการสร้างตัวแบบ ตัวแปรดังกล่าวจะถูก

ตัดหรือปรับปรุงใหม่จนกว่าจะผ่านการคำนวณว่าเป็นคุณลักษณะที่มีค่าความคลาดเคลื่อนที่ยอมรับได้ เพื่อลดมิติของกระบวนการสร้างตัวแบบทางคณิตศาสตร์ (Chirawichitchai, 2018)

การประยุกต์ใช้โปรแกรม RapidMiner Studio เพื่อดำเนินการวิจัยมีขั้นตอนดังนี้

1.1 เลือกใช้ Operation ในขั้นตอนการนำเข้าข้อมูลของผู้ติดเชื้อไวรัสโคโรนา 2019 คือ Retrieve Data และ Automatic Feature Engineering เชื่อมต่อหากันเพื่อหารูปแบบความสัมพันธ์ทางคณิตศาสตร์ โดยตั้งค่าพารามิเตอร์ (ตัวแปร Parameters) ในขั้นตอนวิศวกรรมคุณลักษณะแบบอัตโนมัติ ดัง Table 2 ต่อไปนี้

Table 1 Variables from the personal infection data for the COVID-19

Variable	Definition	Type
x_1	Confirmed	explanatory variable (integer number)
x_2	Deaths	explanatory variable (integer number)
x_3	Recovered	explanatory variable (integer number)
x_4	Active	explanatory variable (integer number)
x_5	New deaths	explanatory variable (integer number)
x_6	New recovered	explanatory variable (integer number)
x_7	Confirmed last week	explanatory variable (integer number)
x_8	Deaths / 100 Cases	explanatory variable (real number)
x_9	Recovered / 100 Cases	explanatory variable (real number)
x_{10}	Deaths / 100 Recovered	explanatory variable (real number)
x_{11}	1 week change	explanatory variable (integer number)
x_{12}	1 week % increase	explanatory variable (real number)
y	New case	responsive variable (integer number)

Table 2 Parameterizing the automatic feature engineering process

Parameter	Set condition
Mode	feature selection
balance for accuracy	1.0
local random seed	1992
maximum generations	30
population size	10
maximum function complexity	20

1.2 ภายใต้ขั้นตอน Automatic Feature Engineering ผู้วิจัยเลือก Operators ที่จะแบ่งข้อมูลด้วย Cross Validation โดยใช้วิธี 10 -fold cross-validation เพื่อเพิ่มประสิทธิภาพของการประเมินรูปแบบที่สร้างตัวแบบให้ดีขึ้น โดยตั้งค่าพารามิเตอร์ดังแสดงใน Table 3

Table 3 Parameterizing the cross validation process

Parameter	Set condition
number of folds	10
sampling type	automatic
use local random seed	selected
local random seed	1992
enable parallel execution	selected

1.3 ภายใต้ขั้นตอน Cross validation ผู้วิจัยเลือก GLM ในการสร้างตัวแบบเชิงเส้นนัยทั่วไป ลงในด้านการฝึกเรียนรู้ Training และผู้วิจัยเลือก Apply Model กับ Performance classification ในด้านการทดสอบการเรียนรู้ โดยตั้งค่าพารามิเตอร์ดังแสดงใน Table 4 ซึ่งการกำหนดพารามิเตอร์ family และ solver เป็น AUTO จะทำให้โปรแกรม RapidMiner เลือก family ของการแจกแจงทางสถิติที่เหมาะสมกับการสร้างตัวแบบและเครื่องมือในการประเมินค่าเหมาะที่สุดอัตโนมัติ

Table 4 Parameterizing the modeling process to select features

Parameter	Set condition
family	AUTO
solver	AUTO
reproducible	selected
maximum number of threads	4
use regularization	selected
standardize	selected
add intercept	selected
missing value handling	meanImputation
max iterations	0
max runtime seconds	0

การประยุกต์ใช้โปรแกรม RapidMiner Studio เพื่อดำเนินการวิศวกรรมคุณลักษณะ มีขั้นตอนสามารถสรุปได้ดังผังแสดงขั้นตอนวิธีใน Figure 1

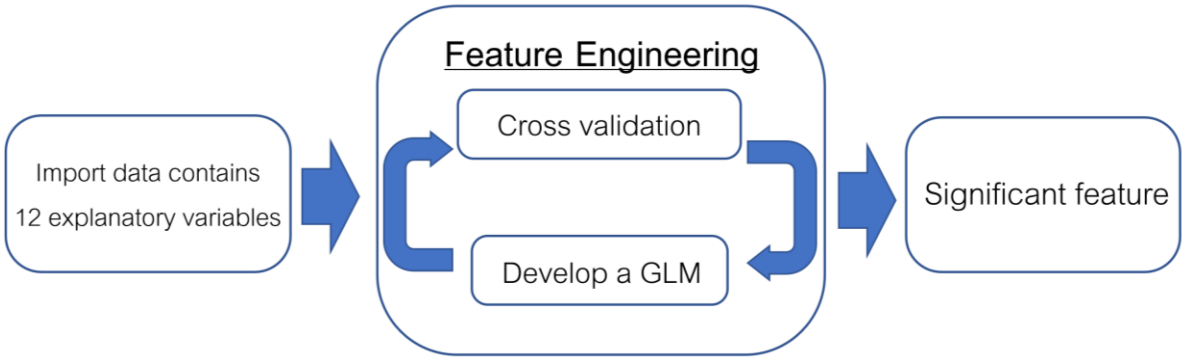

Figure 1 Characterization algorithm for feature screening by RapidMiner Studio program

ขั้นตอนที่ 2 การสร้างตัวแบบเชิงเส้นนัยทั่วไปเพื่อทำนายจำนวนผู้ติดเชื้อไวรัสโคโรนา 2019 ด้วยโปรแกรม RStudio

แนวคิดการประยุกต์ใช้โปรแกรม RStudio เพื่อดำเนินการสร้างตัวแบบเชิงเส้นนัยทั่วไป มีขั้นตอนสามารถสรุปได้ดังผังแสดงขั้นตอนวิธีใน Figure 2

ขั้นตอนการประยุกต์ใช้โปรแกรม RStudio เพื่อดำเนินการสร้างตัวแบบเชิงเส้นนัยทั่วไป มีดังต่อไปนี้

2.1 การนำข้อมูลประเภท CSV มาใช้ในการสร้างตัวแบบดำเนินการโดยนำข้อมูลเข้ามาใส่ตัวแปรที่กำหนด

2.2 การแบ่งข้อมูลเพื่อเรียนรู้และทดสอบการปรับปรุงประสิทธิภาพของตัวแบบในการศึกษานี้ โดยใช้วิธีการสุ่มแบ่งชุดข้อมูลเป็น 2 ส่วน คือ ชุดที่ใช้เรียนรู้เพื่อสร้างตัวแบบและชุดที่ใช้ทดสอบผลการทำนายที่ได้จากตัวแบบ ในอัตราส่วน 70:30

2.3 การสร้างตัวแบบเชิงเส้นนัยทั่วไป

การสร้างตัวแบบเชิงเส้นนัยทั่วไปต้องการพยากรณ์ตัวแปรจำนวนผู้ติดเชื้อรายใหม่ (New case) เพื่อคำนวณค่าคะแนนขององค์ประกอบหลักที่ได้ของแต่ละตัวชี้วัดโดยการหาค่าพารามิเตอร์ปรับค่าที่เหมาะสมสำหรับข้อมูลที่ต้องการจุดประสงค์ของการใช้ตัวแบบเชิงเส้นนัยทั่วไป คือ การหาความสัมพันธ์ระหว่างตัวแปรตอบสนอง (response variable) ซึ่งในงานวิจัยนี้คือจำนวนผู้ติดเชื้อรายใหม่ กับตัวแปรอธิบาย (explanatory variable) ซึ่งได้แก่ตัวแปรอื่น ๆ อีก 12 ตัวแปร

2.4 คัดเลือกตัวแปรแบบลำดับขั้น เพื่อคัดสรรตัวแปรที่มีนัยสำคัญต่อการสร้างตัวแบบ

2.5 ตรวจสอบประสิทธิภาพการพยากรณ์ของตัวแบบเชิงเส้นนัยทั่วไป

โดยรวม ขั้นตอนวิธีดำเนินการวิจัยในครั้งนี้สามารถสรุปรวมเป็นผังภาพได้ ดังแสดงใน Figure 3

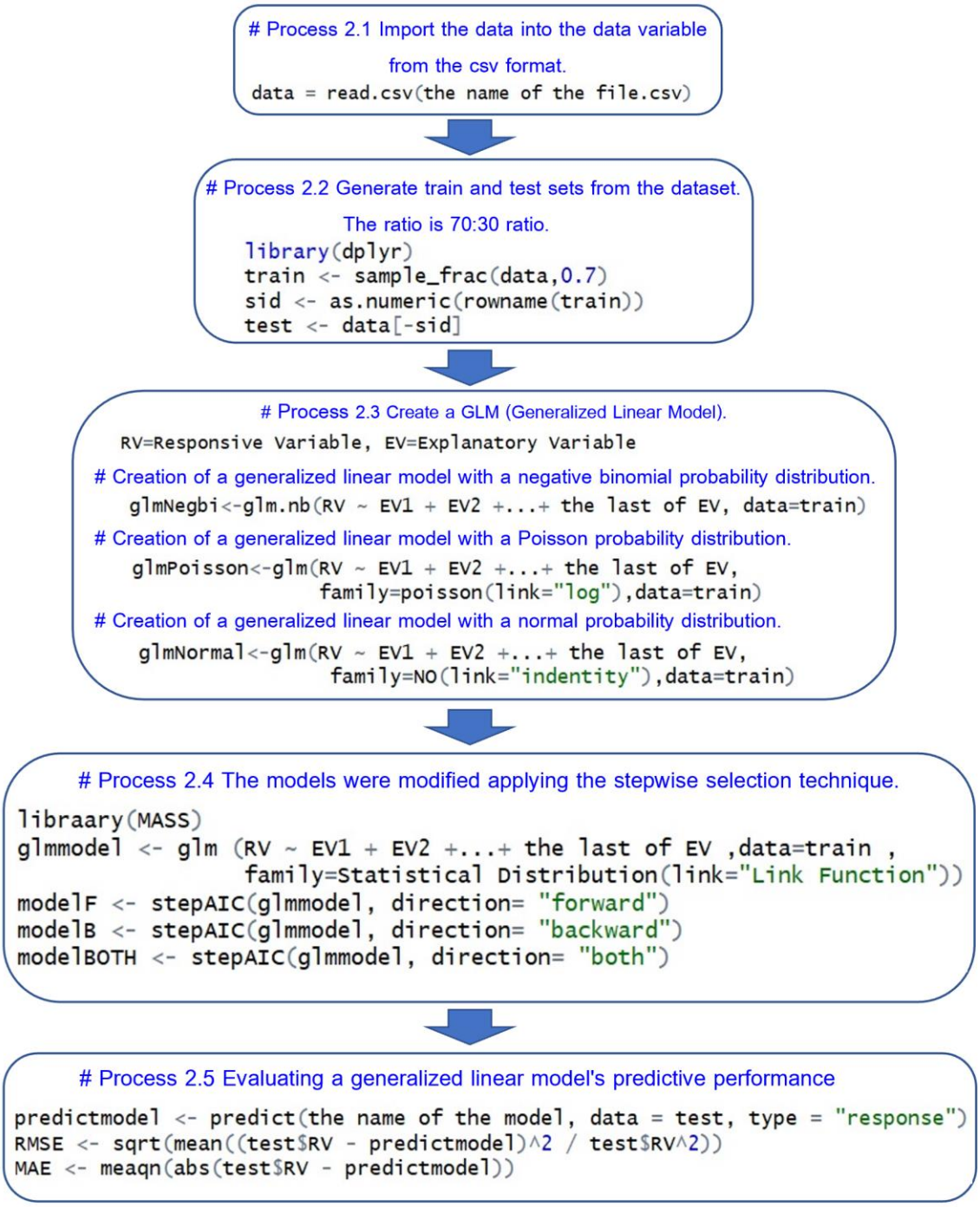

Figure 2 Algorithm for constructing a generalized linear model to forecast the number of COVID-19 new cases and an example set of R language instructions

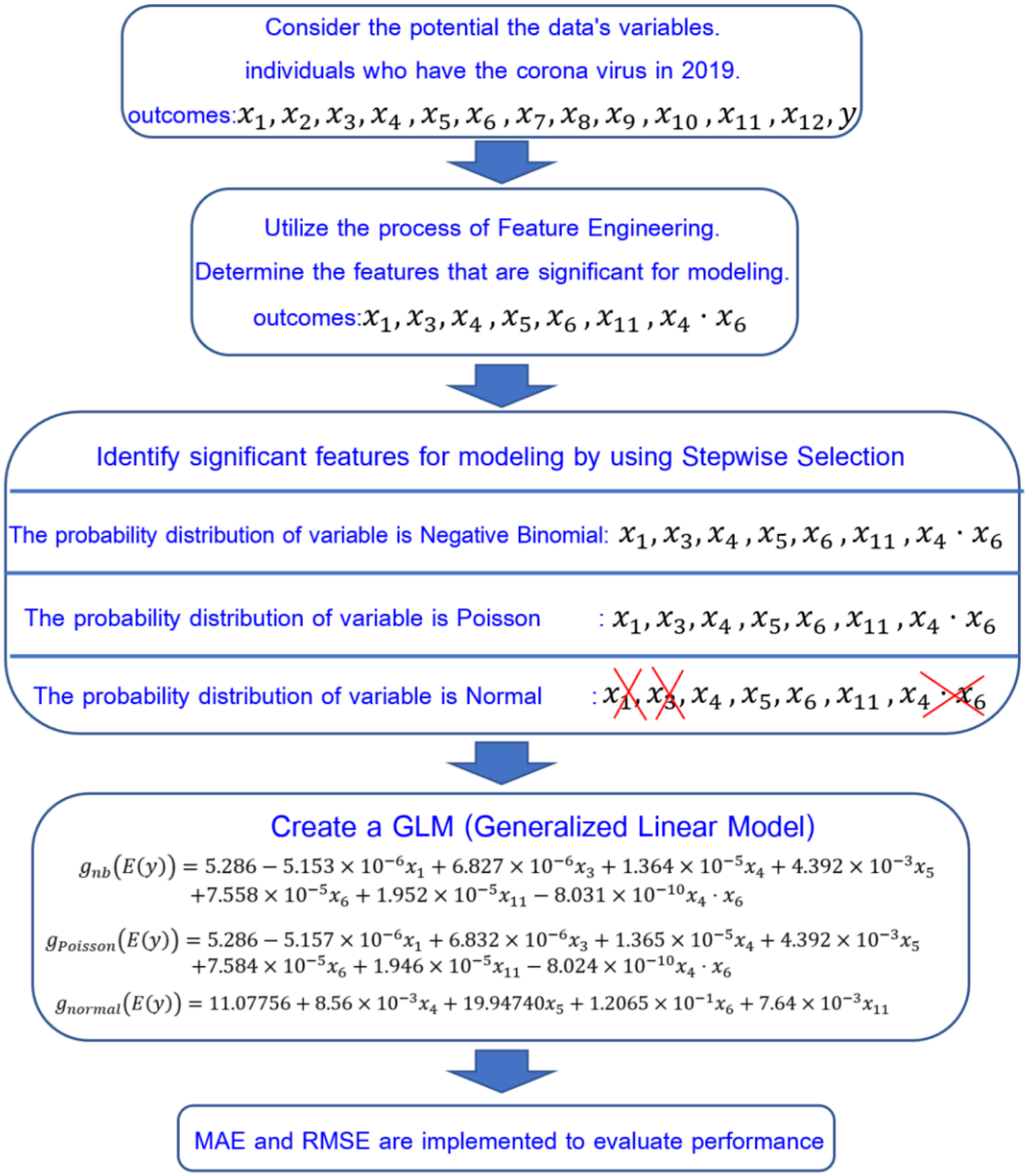

Figure 3 Algorithm for forecasting the number of COVID-19 new cases utilizing feature engineering, stepwise selection, and generalized linear model

ผลการวิจัย

ผลวิเคราะห์โดยโปรแกรม RapidMiner ในการหาคุณลักษณะที่ความสัมพันธ์กับตัวแปรจำนวนผู้ติดเชื้อรายใหม่ (New case) อย่างมีนัยสำคัญ พบว่ามีคุณลักษณะ 7 อย่าง จาก 6 ตัวแปรอธิบาย แสดงดังTable 5 ทั้งนี้พบว่าคุณลักษณะที่ถูกลดทอนออกไปเมื่อทดสอบในการสร้างตัวแบบเบื้องต้นแล้วไม่มีนัยสำคัญต่อการสร้างตัวแบบเชิงเส้นนัยทั่วไปที่กำหนด

Table 5 Features produced by the analysis's application of the feature engineering technique

Feature	Variable	Definition
1	x_1	number of confirmed cases
2	x_3	number of recovered cases
3	x_4	number of active cases
4	x_5	number of new deaths
5	x_6	number of newly recovered cases
6	x_{11}	number of new cases in a week
7	$x_4 \cdot x_6$	product of the number of new recovered cases with the number of active cases

ค่าสัมประสิทธิ์ของคุณลักษณะที่ปรากฏในการสร้างตัวแบบเชิงเส้นนัยทั่วไปที่มีประสิทธิภาพดีที่สุดด้วยโปรแกรม RStudio โดยการประเมิน AIC และการคัดเลือกตัวแปรแบบลำดับขั้น พิจารณาตามการแจกแจงทางสถิติแสดงได้ดัง Table 6 ถึง Table 8 ดังนี้

Table 6 Features selected stepwise selection with negative binomial distributions

Variable	Coefficient	Pr(>|z|)
Intercept	5.286	0.0131
x_1	-0.000005153	0.6983
x_3	0.000006827	0.8594
x_4	0.00001364	0.2377
x_5	0.004392	0.0811
x_6	0.00007558	0.9161
x_{11}	0.00001952	0.9074
$x_4 \cdot x_6$	-0.0000000008031	$<2x10^{-16}$

Table 7 Features selected stepwise selection with Poisson distribution

Variable	Coefficient	Pr(>|z|)
Intercept	5.286	$<2x10^{-16}$
x_1	-0.000005157	$<2x10^{-16}$
x_3	0.000006832	$<2x10^{-16}$
x_4	0.00001365	$<2x10^{-16}$
x_5	0.004392	$<2x10^{-16}$
x_6	0.00007584	$<2x10^{-16}$
x_{11}	0.00001946	$<2x10^{-16}$
$x_4 \cdot x_6$	-0.0000000008024	$<2x10^{-16}$

Table 8 Features selected stepwise selection with normal distribution

Variable	Coefficient	Pr(>|t|)
Intercept	11.07756	0.736
x_4	0.00856	$4.70\text{x}10^{-13}$
x_5	19.94740	$8.75\ \text{x}10^{-13}$
x_6	0.12065	$<2\text{x}10^{-16}$
x_{11}	0.00764	$5.44\text{x}10^{-16}$

จาก Table 6-8 จะได้

สมการตัวแบบเชิงเส้นนัยทั่วไปสำหรับข้อมูลที่มีการแจกแจงทางสถิติแบบทวินามลบ คือ

$$g_{nb}(E(y)) = 5.286 - 5.153 \times 10^{-6}x_1 + 6.827 \times 10^{-6}x_3 + 1.364 \times 10^{-5}x_4 + 4.392 \times 10^{-3}x_5 + 7.558 \times 10^{-5}x_6 + 1.952 \times 10^{-5}x_{11} - 8.031 \times 10^{-10}x_4 \cdot x_6$$

สมการตัวแบบเชิงเส้นนัยทั่วไปสำหรับข้อมูลที่มีการแจกแจงทางสถิติแบบปัวซง คือ

$$g_{Poisson}(E(y)) = 5.286 - 5.157 \times 10^{-6}x_1 + 6.832 \times 10^{-6}x_3 + 1.365 \times 10^{-5}x_4 + 4.392 \times 10^{-3}x_5 + 7.584 \times 10^{-5}x_6 + 1.946 \times 10^{-5}x_{11} - 8.024 \times 10^{-10}x_4 \cdot x_6$$

สมการตัวแบบเชิงเส้นนัยทั่วไปสำหรับข้อมูลที่มีการแจกแจงทางสถิติแบบปรกติ คือ

$$g_{normal}(E(y)) = 11.07756 + 8.56 \times 10^{-3}x_4 + 19.94740x_5 + 1.2065 \times 10^{-1}x_6 + 7.64 \times 10^{-3}x_{11}$$

ผลการวิเคราะห์ประสิทธิภาพด้วย RMSE และ MAE ของตัวแบบทั้ง 3 แสดงดังแผน Diagram 1

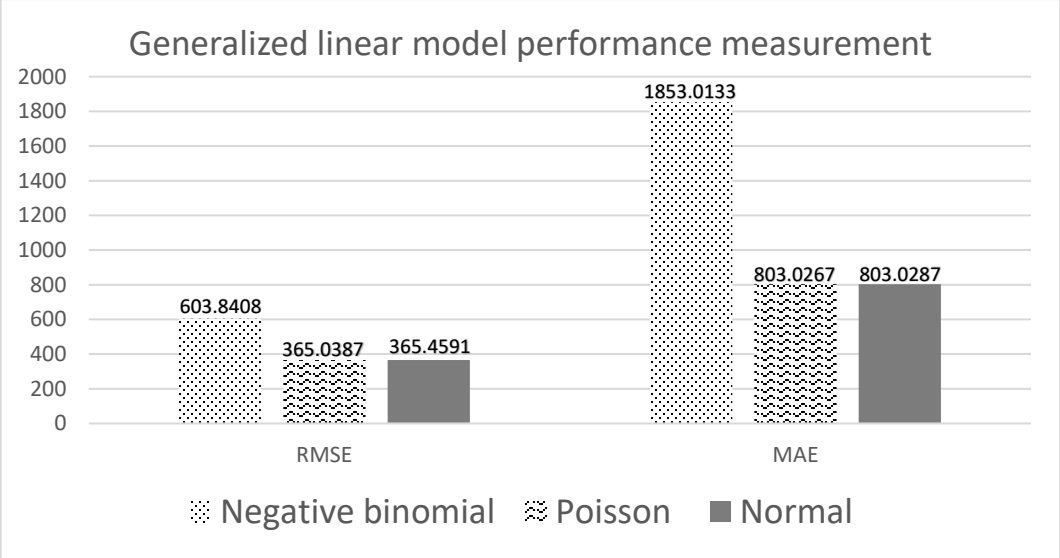

Diagram 1 Display of generalized linear model performance measures

วิจารณ์ผลการวิจัย

ด้วยขั้นตอนวิธีวิศวกรรมคุณลักษณะ ทำให้จำนวนตัวแปรอธิบายลดลงเหลือเพียง 6 ตัวแปร และคุณลักษณะที่เป็นผลคูณของ x_4 และ x_6 ก็ยังคงเป็นคุณลักษณะที่มีนัยสำคัญต่อการสร้างตัวแบบในกรณีพิจารณาว่ามีการแจกแจงความน่าจะเป็นแบบทวินามลบและแบบปัวซง แต่อย่างไรก็ตามสำหรับการสร้างตัวแบบในกรณีพิจารณาว่ามีการแจกแจงความน่าจะเป็นแบบปรกติ การใช้การคัดเลือกตัวแปรแบบลำดับขั้นทำให้คงตัวแปรที่มีนัยสำคัญต่อการสร้างตัวแบบเพียงแค่ 4 ตัวแปรเท่านั้น ซึ่งเมื่อพิจารณา Diagram 1 พบว่า RMSE และ MAE ที่น้อยที่สุด คือ 365.0387 และ 803.0267 ตามลำดับ ซึ่งเป็นค่าของตัวแบบที่มีการแจกแจงทางสถิติแบบปัวซง แสดงว่าเป็นตัวแบบที่มีประสิทธิภาพสูงสุด ตัวแบบที่มีประสิทธิภาพรองลงมาเป็นของตัวแบบที่มีการแจกแจงทางสถิติแบบปรกติโดยมี RMSE และ MAE คือ 365.4591 และ 803.0286 ตามลำดับ และสุดท้ายตัวแบบที่มีการแจกแจงทางสถิติเป็นแบบทวินามลบมีประสิทธิภาพด้อยสุด แต่อย่างไรก็ตามเมื่อเทียบระหว่างตัวแบบที่มีการแจกแจงทางสถิติแบบปัวซงและแบบปรกติ ค่า RMSE และ MAE มีค่าต่างกันเพียงเล็กน้อย แต่ตัวแบบที่มีการแจกแจงทางสถิติแบบปัวซงต้องใช้คุณลักษณะถึง 7 อย่างจาก 6 ตัวแปรอธิบาย ในการสร้างตัวแบบ โดยตัวแบบที่มีการแจกแจงทางสถิติแบบปรกติ ใช้เพียง 4 คุณลักษณะ (จาก 4 ตัวแปรอธิบาย) เท่านั้น ซึ่งมีผลในทางด้านการประมวลผลที่น้อยกว่า ช่วยลดเวลาทั้งทางด้านการสร้างตัวแบบและการพยากรณ์ ทั้งนี้เพื่อทดสอบความสามารถในการพยากรณ์โดยใช้ข้อมูลเกี่ยวกับผู้เชื้อไวรัสโคโรนา 2019 ของประเทศจีน ซึ่งเป็นประเทศที่พบเชื้อไวรัสโคโรนา 2019 เป็นประเทศแรก จากฐานข้อมูลที่ดำเนินการวิจัยมาประมวลผล พบว่า $x_1 = 86{,}783, x_3 = 78{,}869, x_4 = 3{,}258, x_5 = 4, x_6 = 7, x_{11} = 1{,}161$ และ $y = 213$ เมื่อใช้สมการตัวแบบเชิงเส้นนัยทั่วไปสำหรับข้อมูลที่มีการแจกแจงทางสถิติแบบปัวซงจะได้ค่าพยากรณ์คือ 5.46 ในขณะที่ $g_{Poisson}(E(y))$ คำนวณจากฟังก์ชันเชื่อมโยง (link function) ลอการิทึม ทำให้ได้ว่า $\ln y = \ln 213 = 5.36$ ซึ่งมีค่าใกล้เคียงกับค่าพยากรณ์ ขณะที่หากพิจารณาเป็นตัวแบบที่มีการแจกแจงทางสถิติแบบปรกติจะได้ค่าพยากรณ์คือ 120.49 โดยค่า $g_{normal}(E(y))$ คือ 213 (ฟังก์ชันเชื่อมโยงคือฟังก์ชันเอกลักษณ์) ซึ่งค่าพยากรณ์มีความแตกต่างจากค่าจริง เห็นได้ว่าผลที่ได้นี้สอดคล้องกับงานวิจัยของ Xie & Farrell (2020) และ Benti (2022) ซึ่งกล่าวว่าจำนวนผู้ติดเชื้อไวรัสโคโรนา 2019 และการเสียชีวิตของผู้ป่วยที่ติดเชื้อดังกล่าวมีรูปแบบการแจกแจงทางสถิติแบบปัวซง

สรุปผลการวิจัย

งานวิจัยครั้งนี้ได้พัฒนารูปแบบของการประยุกต์ใช้วิศวกรรมคุณลักษณะร่วมกับการสร้างตัวแบบเชิงเส้นนัยทั่วไปเพื่อพยากรณ์จำนวนผู้ติดเชื้อไวรัสโคโรนา 2019 รายใหม่ โดยใช้ข้อมูลจากแหล่งข้อมูลสาธารณะ และเป็นข้อมูลจากหลายหลายประเทศ โดยพิจารณาการสร้างตัวแบบโดยใช้ในรูปแบบการแจกแจงทางสถิติ 3 อย่างได้แก่ การแจกแจงแบบทวินามลบ การแจกแจงแบบปัวซง และการแจกแจงแบบปรกติ ทั้งนี้ผลการวิจัยชี้ให้เห็นว่าจากข้อมูลที่ใช้ในการวิจัยซึ่งมีตัวแปรอธิบายจำนวน 13 ตัวแปร เมื่อใช้วิศวกรรมคุณลักษณะเข้ามาดำเนินการ พบว่าตัวแปรที่มีนัยสำคัญในการสร้างตัวแบบจะเหลือเพียง 6 ตัวแปร และสามารถใช้คุณลักษณะเพียง 7 คุณลักษณะเท่านั้นในการสร้างตัวแบบ นอกจากนี้ถึงแม้การสร้างตัวแบบเพื่อพยากรณ์ด้วยการแจกแจงแบบปัวซงจะให้ผลการประเมินประสิทธิภาพที่ดีที่สุดแต่ก็มีประสิทธิภาพดีกว่าตัวแบบที่ใช้การแจกแจงแบบปรกติเพียงเล็กน้อย ซึ่งตัวแบบที่ใช้การแจกแจงแบบปรกติกลับใช้คุณลักษณะในการพยากรณ์ที่น้อยกว่าพอสมควร โดยใช้

เพียง 4 คุณลักษณะ ขณะที่การแจกแจงแบบปัวซงจะใช้ถึง 7 คุณลักษณะ ซึ่งมีประโยชน์ให้การวางแผนในการจัดเก็บข้อมูลเพื่อใช้ในการพยากรณ์ต่อในอนาคตมีความสะดวกมากขึ้นเพราะจัดเก็บข้อมูลน้อยลง ทำให้ประหยัดเวลาในการจัดเก็บข้อมูลและประมวลผลเพื่อรับมือสถานการณ์การระบาดของโรค ทั้งนี้ส่งผลต่อความรวดเร็วในการดำเนินการก้าวผ่านวิกฤตเพื่อกลับสู่ภาวะปกติในการชีวิตต่อไป

กิตติกรรมประกาศ

คณะผู้ทำวิจัยได้รับทุนสนับสนุนจากทุนในโครงการพัฒนาและส่งเสริมผู้มีความสามารถทางด้านวิทยาศาสตร์และเทคโนโลยี (พสวท.) และทุนวิจัยจากแหล่งทุนภายนอกจากกองทุนสนับสนุนการวิจัยและพัฒนา (ทุนOROG) มหาวิทยาลัยเทคโนโลยีสุรนารี งานวิจัยนี้สำเร็จสมบูรณ์ได้ด้วยความร่วมมือของบุคลากรสาขาคณิตศาสตร์ และ สาขาวิศวกรรมนวัตกรรมชีวการแพทย์สำนักวิชาวิศวกรรมศาสตร์หลายท่าน ผู้จัดทำขอขอบพระคุณไว้ ณ ที่นี้

เอกสารอ้างอิง

Amattayakul, S. (2020). *The world after COVID-19 economic and social impact*. Strategy and Planning Division Foreign Commerce Group, Office of the Permanent Secretary, Ministry of Commerce. Interview on 2020/7/16 (in Thai)

Benlagha, N. (2020). *Modeling the Declared New Cases of COVID-19 Trend Using Advanced Statistical Approaches, Preprint Document.* DOI:10.6084/m9.figshare.12052638.

Benti, T. B. (2022). *Modeling Mortality from COVID-19 Using Poisson Based Regressions: The Case of Sweden.* U.U.D.M. Project Report 2022:9, Department of Mathematics, Uppsala University.

Chirawichitchai, H. (2018). *AutoFE: Efficient and Robust Automated Feature Engineering.* Master of Engineering in Electrical Engineering and Computer Science at the Massachusetts Institute of Technology, 13-15.

Department of Mental Health. (2022). What does endemic disease mean and why is it in the category of coronavirus? *News from newspapers related to mental health*. (in Thai)

Emerging Infectious Disease Work of Communicable Disease Academic Development Group. (2021). *Coronavirus disease 2019 (COVID-19) situation, public health measures and barriers to disease prevention and control among travelers.* Department of Disease Control. (in Thai)

javaTpoint. (2021). *K-Nearest Neighbor (KNN) Algorithm for Machine Learning*.

Kasilingam, D., Prabhakaran, S.P.S., Rajendran, D.K. Rajagopal, V., Kumar, T.S., & Soundararaj, A. (2021). Exploring the growth of COVID-19 cases using exponential modelling across 42 countries and predicting signs of early containment using machine learning. *Transboundary and Emerging Diseases. 86.* Wiley Online Library.68(3), 1001-1018.

Kelter, D., Ghiassi, K., Patel, S., Connors, C., Bonk, M. P., Gray, E., Zarbiv, S. A., Menon, A., & Juneja, P. (2021). Use of feature engineering to predict COVID-19 mortality. *American Thoracic Society International Conference Abstracts. American Journal of Respiratory and Critical Care Medicine 2021, 203, A2630*

Leelarutsamee, A. (n.d.). *Interesting Facts about COVID-19 Infection from SARS-CoV-2*. The Medical Council of Thailand. (in Thai)

Nawaratana N. (2019). *Analysis of distributions for insurance claims data.* Master Degree Thesis of Suranaree University of Technology, 37–38.

Office of the Royal Thai Embassy. (2018). *Dictionary of Statistical Terms, Royal Thai Council edition*. 2nd ed. (amended). Chulalongkorn University Press. (in Thai)

Patcharawongsakda, A. (2014). *Introduction to Data Analysis with Data Mining Techniques*. Bangkok: Asia Digital Press Company Limited. (in Thai)

Reis, G. F. N. (2019). Automated Feature Engineering for Classification Problems. *American Faculdade De Engenharia Da Universidade Do Porto*. 5-11.

Strategy and Organization Development Group. (2022). *Government Action Plan for the Fiscal Year 2022*. Urban Disease Prevention and Control Institute, Ministry of Public Health. (in Thai)

Vytla1, V., Ramakuri, S.K., Peddi, A., Srinivas, K.K. & Ragav, N.N. (2021*)*. Mathematical Models for Predicting Covid-19. *Journal of Physics: Conference Series, 1797(2001)012009,* Doi: 10.1088/1742-6596/1797/1/012009.

Xie, J. & Farrell, P. (2020). *Analysis of COVID-19 Confirmed Cases based on Poisson Loglinear Regression Model.* Honours Project. School of Mathematics and Statistics. Carleton University.