

อัลกอริธึมการเรียนรู้ของเครื่องสำหรับการทำนายการคัดแยกผู้ป่วย COVID-19

Towards Machine Learning Algorithm for Screening Prediction of COVID-19 Patients

ศิวกกร บรรลือทรัพย์* และ วราพร จิระพันธุ์ทอง

Siwakorn Banluesapy and Waraporn Jirapanthong*

สาขาวิชาวิศวกรรมเว็บและการพัฒนาแอปพลิเคชันบนอุปกรณ์พกพา

วิทยาลัยครีเอทีฟดีไซน์แอนด์เอ็นเตอร์เทนเมนต์เทคโนโลยี มหาวิทยาลัยธุรกิจบัณฑิต

Department of Web Engineering and Mobile Application Development

College of Creative Design and Entertainment Technology Dhurakij Pundit University

Received: December 27, 2021; Revised: April 11, 2022; Accepted: June 07, 2022; Published: June 29, 2022

ABSTRACT – This research aims to develop a web-based application for screening COVID-19 patients by applying with machine learning techniques. We have applied the techniques to classify the data and then compare the algorithm's efficiency in prediction of COVID-19 patient screening. The data on 1,608,923 patients from the Department of Disease Control are accumulated. In particular, the data from January 1, 2020 to October 1, 2021 are learnt by three classification algorithms: Random Forest, Neural network, and Naive Bayes. The performance of predictive techniques between features are compared. The models learnt and generated based on different three algorithms are tested by the Split Test method. The data are randomly divided into two parts: Training Data (80%) and Test Data (20%) by using Orange Canvas. The models are then experimented to determine the performance test results with the highest accuracy. As the results, it is found that the accuracy rates of Random Forest, Neural Network, and Naive Bayes are 93.33%, 92.7%, and 92.1%, respectively. The model based on Random Forest technique with the highest value of accuracy is then applied in the application for screening COVID-19 patients.

KEYWORDS: Random forest, Neural network, Naive Bayes, Machine Learning, Covid-19

บทคัดย่อ -- งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาระบบทำการคัดแยกผู้ป่วยโควิด-19 ในรูปแบบเว็บแอปพลิเคชันโดยใช้เทคนิคการเรียนรู้ของเครื่องในการจำแนกข้อมูลจากนั้นทำการเปรียบเทียบประสิทธิภาพของอัลกอริทึมในการทำนายการคัดแยกผู้ป่วยโควิด-19 และได้ทำการศึกษาข้อมูลผู้ป่วย จำนวน 1,608,923 ราย จากกรมควบคุมโรค ทำการรวบรวมข้อมูล 1 มกราคม 2563 ถึง วันที่ 1 ตุลาคม 2564 โดยใช้อัลกอริทึมการจำแนกข้อมูลทั้งหมด 3 แบบได้แก่ Random Forest, Neural network และ Naive Bayes ซึ่งทำการเปรียบเทียบประสิทธิภาพรูปแบบเทคนิคทำนายระหว่างคุณลักษณะ การทดสอบประสิทธิภาพตัวแบบทำนายด้วยวิธีการ Split Test คือ การแบ่งข้อมูลด้วยการสุ่มออกเป็น 2 ส่วนคือ Training Data (ร้อยละ 80) และ Test Data (ร้อยละ 20) โดยใช้โปรแกรม Orange Canvas จากนั้นทำการทดลองเพื่อหาผลการทดสอบประสิทธิภาพที่มีค่าความถูกต้องที่สูงที่สุด ผลการวิจัยพบว่า Random Forest มีค่าความถูกต้องเท่ากับ 93.33% Neural

*Corresponding Author: siwakorn167@gmail.com

network มีค่าความถูกต้องเท่ากับ 92.7% Naive Bayes มีค่าความถูกต้องเท่ากับ 92.1% จากผลการเปรียบเทียบในครั้งนี้จึงนำ Random Forest ที่มีค่าสูงที่สุดไปใช้ในการพัฒนาโปรแกรมระบบในการคัดแยกผู้ป่วยโควิด-19 ได้

คำสำคัญ: ป่าลุ่ม, โครงข่ายประสาทเทียม, นาอ์เบย์, การเรียนรู้ของเครื่อง, โควิด-19

1. บทนำ

โรคโควิด-19 เป็นโรคติดต่อสายพันธุ์ใหม่ และได้มีการคาดการณ์ว่าเป็นการติดเชื้อจากสัตว์สู่คนโดยมีค้างคาวเป็นพาหะเป็นโรคได้ทั้งคนและสัตว์ [1], [2] เชื้อสามารถติดต่อจากคนสู่คนผ่านทางเดินหายใจจากละอองของเสมหะจากการ ไอ จาม น้ำมูก น้ำลายในระยะประมาณ 1-2 เมตร [3]

สำหรับการระบาดของเชื้อไวรัสโควิด-19 รอบแรกในประเทศไทย พบผู้ป่วยต้องสงสัยรายแรกเมื่อวันที่ 21 มกราคม พ.ศ. 2563 เป็นนักท่องเที่ยวหญิงชาวจีนอายุ 74 ปี ซึ่งเดินทางมาถึงกรุงเทพมหานครโดยเที่ยวบินจากนครอู่ฮั่น เมืองหลวงของมณฑลหูเป่ย์ สาธารณรัฐประชาชนจีน และ ณ วันที่ 8 เมษายน พ.ศ. 2563 พบอัตราการติดเชื้อในประเทศไทย จำนวน 2,369 คน [4] หลังจากนั้นพบการระบาดระลอกใหม่ช่วงปลายปี 2563 เริ่มตั้งแต่วันที่ 17 ธันวาคม 2563 พบหญิงไทย อายุ 67 ปี อาชีพค้าขายที่ตลาดกลางกุ้ง ในตำบลมหาชัย อำเภอมหาชัย จังหวัดสมุทรสาคร [5]

กระบวนการตรวจ วินิจฉัยโรคโควิด 19 เป็นการตรวจทางห้องปฏิบัติการ ต้องเสียเวลาในการเตรียมเชื้อและกระบวนการตรวจ เพราะต้องตรวจกับตัวเชื้อจริง ในช่วงแรกของการระบาดต้องใช้เวลานานไม่น้อยกว่า 2 วัน และทำได้เฉพาะในโรงพยาบาลใหญ่หรือโรงพยาบาลศูนย์ ค่าตรวจมีราคาแพง เพราะผลที่ได้ต้องได้ผลชัดเจน และ ทางกระทรวงสาธารณสุขมีข้อจำกัดในเรื่องจำนวนการตรวจ จึงต้องคัดกรองผู้ต้องสงสัยก่อนเก็บเชื้อส่งตรวจ และใช้วิธีนี้ตั้งแต่เริ่มการระบาด คือ ต้องเก็บเชื้อจากสารคัดหลั่งที่โพรงจมูก ดังนั้นในกระบวนการควบคุมโรค การค้นหา การตรวจหาผู้ป่วย จำเป็นต้องมีกระบวนการคัดกรองเพื่อคัดกรองเป็นลำดับในการหาผู้สงสัยก่อน [6] ดังนั้นผู้วิจัยจึงทำการศึกษาและวิเคราะห์ข้อมูลผู้ติดเชื้อ โควิด-19 จากกรมควบคุมโรคตั้งแต่ระลอกที่ 1 – 3 ซึ่งเป็นข้อมูลแบบเปิด โดยทำการรวบรวมข้อมูลตั้งแต่วันที่ 1 มกราคม 2563 ถึง วันที่ 1 ตุลาคม 2564 เป็นจำนวน 1,608,923 ราย [7] เพื่อช่วยคัดแยกประเภทผู้ป่วยได้เร็วขึ้น

การเรียนรู้ของเครื่อง (Machine Learning) [8] คือ การที่สามารถทำให้คอมพิวเตอร์สามารถเรียนรู้ได้ด้วยตนเอง โดยการเขียนโปรแกรมหรือการสร้างอัลกอริทึมที่สามารถเรียนรู้และพยากรณ์ข้อมูลได้ ทั้งนี้อัลกอริทึมจะต้องอาศัยโมเดลที่ได้จากการสร้างชุดของข้อมูลเพื่อพยากรณ์หรือตัดสินใจ

งานวิจัยนี้จึงใช้เทคนิคการเรียนรู้ของเครื่องมาช่วยในดำเนินการ ซึ่งเป็นงานในลักษณะ Classification อยู่ในรูปแบบโมเดลประเภท Supervised ซึ่งมีตัวแปรที่ใช้ในการวัดเป้าหมายโดยการคัดแยกประเภทผู้ป่วย เพื่อคัดกรองผู้ต้องสงสัยได้เร็วขึ้น และได้เตรียมการรักษาได้ทันเวลา

2. งานวิจัยและทฤษฎีที่เกี่ยวข้อง

ระบบคัดแยกผู้ป่วยโควิด-19 ด้วยวิธีการนำการเรียนรู้ของเครื่องมาประยุกต์ใช้ในการทำงานวิจัยฉบับนี้ โดยใช้ข้อมูลจากกรมควบคุมโรค [7] ซึ่งรวบรวมตั้งแต่ 1 มกราคม 2563 ถึง วันที่ 1 ตุลาคม 2564 เป็นจำนวน 1,608,923 ราย จากเว็บไซต์ <https://covid19.ddc.moph.go.th/> หรือ DDC Open data

2.1 การประยุกต์ใช้การเรียนรู้ของเครื่องสำหรับโรคโควิด-19

ในช่วงเร่งด่วนวาระระดับโลกเมื่อเร็วๆ นี้ มีนักวิทยาศาสตร์ แพทย์ และผู้เชี่ยวชาญด้านการดูแลสุขภาพทั่วโลกยังคงค้นหาเทคโนโลยีใหม่ๆ เพื่อสนับสนุนการแก้ปัญหาการระบาดใหญ่ของโควิด-19 [9] การเรียนรู้ของเครื่อง (machine learning) ก็เป็นเทคโนโลยีที่เข้ามาช่วยในการแก้ปัญหาของโรคโควิด-19 เช่น จีนใช้การวิเคราะห์ข้อมูลขนาดใหญ่ (Big data analytics) เข้ามาควบคุมการแพร่ระบาดให้เกิดประสิทธิผลและเพื่อให้การดำเนินธุรกิจเป็นปกติมากที่สุด

ปัจจุบันมีการพัฒนาอย่างต่อเนื่องในเทคโนโลยีการเรียนรู้ของเครื่องและปัญญาประดิษฐ์ในงานการแพร่ระบาดของโรคโควิด-19 เพื่อวิเคราะห์การใช้ยา การคัดกรอง การคาดการณ์ การติดตามผู้สัมผัส ตลอดจนกระบวนการพัฒนาวัคซีนของโรคโควิด-19 อย่างมีนัยสำคัญ เพื่อลดการแทรกแซงของมนุษย์

ในการปฏิบัติการทางการแพทย์ แต่อย่างไรก็ตามโมเดลส่วนใหญ่ไม่ได้ถูกนำไปใช้งานมากพอที่จะสามารถแสดงการปฏิบัติงานในโลกแห่งความเป็นจริง แต่ก็ยังสามารถใช้รับมือการแพร่ระบาดของโรคโควิด-19 ได้ [10]

2.2 สถานการณ์ผู้ติดเชื้อ COVID-19 อพเททรายวัน

รูปแบบ API (Json/CSV Data Format)

สถานการณ์ผู้ติดเชื้อ COVID-19 อพเททรายวัน รูปแบบ API (Json/CSV Data Format) เป็นข้อมูลแบบเปิดที่รวบรวมโดยกรมควบคุมโรคจากกระทรวงสาธารณสุข สามารถบอกรายละเอียดเป็นพจนานุกรมข้อมูล (Data dictionary) ได้ดังตารางที่ 1

ตารางที่ 1. พจนานุกรมข้อมูลข้อมูลผู้ป่วยโควิด-19

Colum Name	Data Type	Comment
txn_date	date	วันแถลง
gender	varchar	เพศ
age_number	varchar	อายุ
age_range	varchar	ช่วงอายุ
nationality	varchar	สัญชาติ
job	varchar	อาชีพ
risk	varchar	ปัจจัยเสี่ยง
patient_type	text	ประเภทผู้ป่วย
province	varchar	จังหวัด
update_date	datetime	วันที่ปรับปรุงข้อมูล

ตารางที่ 2: Class ในการวิจัย [11], [12]

ประเภทของผู้ป่วย	
Class	Description
1	การเฝ้าระวังในกลุ่มผู้ป่วยหรือผู้มีอาการเข้าได้กับนิยามผู้สงสัยติดเชื้อไวรัสโคโรนา 2019 ที่เข้าเกณฑ์สอบสวนโรค (patient under investigation : PUI)
2	การตรวจคัดกรองในประชากรเสี่ยงตามจุดคัดกรองและด่านเข้าออกระหว่างประเทศ (screening)
3	การเฝ้าระวังในกลุ่มเป้าหมายเฉพาะหรือพื้นที่เฉพาะ (sentinel surveillance)
4	การเฝ้าระวังเหตุการณ์ในสถานที่เสี่ยง เก็บตัวอย่างส่งตรวจเมื่อเข้านิยาม PUI หรือเป็นกลุ่มก้อนของผู้ป่วยทางเดินหายใจ

2.3 การเรียนรู้ของเครื่อง (Machine Learning)

การเรียนรู้ของเครื่อง (Machine Learning) เป็นหนึ่งในประเภทของปัญญาประดิษฐ์ ที่จะช่วยให้แอปพลิเคชัน หรือซอฟต์แวร์เกิดการเรียนรู้จากชุดข้อมูลที่สามารถทำนายผลได้อย่างแม่นยำ โดยหลักการคือจะมีโมเดลทางคณิตศาสตร์และมีอัลกอริทึมในหลากหลายรูปแบบมาเรียนรู้จากชุดข้อมูล (Data set) จนทำให้คอมพิวเตอร์เกิดการเรียนรู้และฉลาดในการตัดสินใจบนข้อมูลที่ไม่เคยได้เห็นมาก่อนได้สำเร็จ

2.4 Random Forest

หลักการของ Random Forest คือการสร้าง model จาก Decision Tree หลายๆ model ย่อยๆ (ตั้งแต่ 10 model ถึง มากกว่า 1000 model) โดยแต่ละ model จะได้รับ data set ไม่เหมือนกัน ซึ่งเป็น subset ของ data set ทั้งหมด ตอนทำ prediction จะให้ Decision Tree ทำการ prediction ในแต่ละโหนด และคำนวณผล prediction ด้วยการ vote output ที่ ถูกเลือกโดย Decision Tree มากที่สุด [13]

2.5 Naive Bayes

เป็นอีกหนึ่งวิธีที่ได้รับความนิยมเนื่องจากการโมเดลง่ายและไม่ซับซ้อน โดยจะอาศัยทฤษฎีความน่าจะเป็น (probability) เป็นหลักซึ่งถูกใช้ในการทำนายผลจัดเป็นเทคนิคในการแก้ปัญหาแบบ Classification ที่สามารถคาดการณ์ผลลัพธ์ได้และสามารถอธิบายได้ โดยจะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปร เพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ [14]

2.6 Neural network

ลักษณะทั่วไปของโครงข่ายประสาทเทียม คือ โหนดจำลองมาจากรอยประสานประสาท (Synapse) ระหว่างใยประสาทนำเข้า (Dendrite) ของเซลล์ประสาทตัวหนึ่งและแกนประสาทนำออก (Axon) ของเซลล์ประสาทอีกตัวหนึ่ง โดยมีฟังก์ชันการแปลงเป็นตัวกำหนดสัญญาณส่งออก องค์ประกอบในการทำงานของระบบประสาทในสมองมนุษย์กับโครงข่ายประสาทเทียมมีความคล้ายคลึงกัน ในระบบประสาทของมนุษย์การรับกระแสประสาทจะอาศัยเดนไดรต์ (Dendrites) เป็นตัวนำพาเปรียบเทียบกับหน่วยข้อมูลป้อนเข้าในระบบโครงข่าย

ประสาทเทียม ส่วนการนำกระแสประสาทไปยังเซลล์ประสาทอื่นผ่านทางเอกซอน ที่เป็นตัวส่งออกกระแสประสาทเปรียบเทียบกับหน่วยผลลัพท์ในโครงข่ายประสาทเทียม [15]

2.7 ตัววัดประสิทธิภาพของโมเดลการจำแนกประเภทข้อมูล (Confusion matrix)

การทำงานของโมเดลไปใช้งานจริงนั้นจะเป็นจะต้องทราบประสิทธิภาพของโมเดลเสียก่อน โดยทั่วไปแล้วจะมีตัววัดที่นิยมใช้กันในงานวิจัยและการทำงานต่างๆ อยู่ 4 ค่า [16], [17] คือ

1. Precision เป็นการวัดความแม่นยำของโมเดล โดยพิจารณาแยกทีละ Class
2. Recall เป็นการวัดความถูกต้องของโมเดล โดยพิจารณาแยกทีละ Class
3. F-measure เป็นการวัดค่า Precision และ Recall พร้อมกันของโมเดลโดยพิจารณาแยกทีละ Class
4. Accuracy เป็นการวัดความถูกต้องของโมเดลโดยพิจารณา รวมทุก Class

แนวคิดเกี่ยวกับ Confusion matrix จากบทความ [14] ได้กล่าวไว้ว่า Confusion matrix คือตารางแบบจัตุรัสโดยมีจำนวนแถวเท่ากับจำนวน Class ดังตัวอย่างในตารางที่ 2 มี Class คำตอบอยู่ 2 คำตอบ คือ yes และ no

ตารางที่ 3. ตัวอย่างแสดงตาราง Confusion matrix

Predicted / actual	Yes	No
Yes	TP	FP
No	FN	TN

จากตารางที่ 3 ค่าที่แสดงในช่องต่างๆ ประกอบไปด้วย

- True Positive (TP) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็น Class ซึ่งกำลังสนใจอยู่
- True Negative (TN) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็น Class ซึ่งไม่ได้สนใจอยู่
- False Positive (FP) คือ จำนวนข้อมูลที่ทำนายผิดมาเป็น Class ซึ่งกำลังสนใจอยู่
- False Negative (FN) คือ จำนวนข้อมูลที่ทำนายผิดมาเป็น Class ซึ่งกำลังสนใจอยู่

2.8 งานวิจัยที่เกี่ยวข้อง

Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks [18] ในงานวิจัยนี้ได้ทำการศึกษาจากผู้ป่วย 108 รายที่เป็นโรคปอดอักเสบจากโควิด-19 จากห้องปฏิบัติการ โดยเป็นชาย 60 คน หญิง 48 คน มีอายุเฉลี่ย 61.45 ± 15.04 ทั้งนี้ชุดข้อมูลได้รวบรวมตั้งแต่ กันยายน 2563 - ธันวาคม 2563 จากโรงพยาบาลในมหาวิทยาลัยที่วิจัย โดยการใช้เทคนิค Deep learning ในการจำแนกคุณลักษณะจากภาพเอกซเรย์คอมพิวเตอร์ เพื่อหาความแตกต่างในการจำแนกผู้ป่วยที่เป็นโควิด-และไม่เป็นโควิด ซึ่งโมเดลที่ใช้จะเป็น ResNet-101 เป็น Deep Convolutional Neural Networks โดยค่า Accuracy อยู่ที่ 99.51% และ Specificity อยู่ที่ 99.0%

Automated detection of COVID-19 cases using deep neural networks with X-ray images [19] งานวิจัยนี้ได้ประยุกต์ใช้การเรียนรู้ของเครื่อง (Machine Learning) ในการทำงานแบบอัตโนมัติ ทั้งนี้ยังได้ใช้ Deep learning ซึ่งเป็นเทคนิคยอดนิยมของปัญญาประดิษฐ์ (AI) โดยชุดข้อมูลเป็นภาพเอกซเรย์ทรวงอก ภาพรังสีวิทยา เป็นชุดข้อมูลแบบเปิดจากเว็บไซต์ <https://github.com/ieee8023/COVID-chestxray-dataset> โดยจะทำการศึกษาการจำแนกผู้ป่วยโควิด, โรคปอดอักเสบ และปอดปกติ จากภาพเอกซเรย์ทรวงอก เพื่อลดภาระงานของแพทย์ในการคัดกรองผู้ป่วยและช่วยประหยัดเวลามากขึ้น โดยค่า Accuracy อยู่ที่ 98.08% จากโมเดล neural network

Combination of four clinical indicators predicts the severe/critical symptom of patients infected COVID-19 [20, p. 19] ได้ทำการวิจัยและศึกษาจากตัวอย่างข้อมูลผู้ป่วยทั้งหมด 336 รายที่ติดเชื้อโควิด-19 ในเซี่ยงไฮ้ ประเทศจีน นอกจากนี้ยังรวบรวมการสังเกต / บันทึกทางคลินิกและทางห้องปฏิบัติการ 220 รายการ ตัวบ่งชี้ทางคลินิกสัมพันธ์กับอาการรุนแรง / อาการวิกฤต ได้มีการระบุ และได้มีการพัฒนาแบบจำลองสำหรับการทำนายอาการรุนแรง / วิกฤต โดยใช้ Support Vector Machine (SVM) โดยให้ค่า Accuracy อยู่ที่ 77.5% และ Specificity อยู่ที่ 78.4%

Covid-19 Classification Using Deep Learning in Chest X-Ray Images [21] ได้ทำการวิจัยข้อมูลผู้ป่วยโรคโควิด-19 จากชุดข้อมูลภาพเอกซเรย์ทรวงอก ภาพรังสีวิทยา เป็นชุดข้อมูลแบบเปิด โดยได้ใช้เทคนิค Deep learning ในรูปแบบโมเดล

ResNet-18 (Residual Network) ซึ่งเป็น Deep Convolutional Neural Networks ที่ทาง Microsoft เผยแพร่แก่ผู้สาธารณสุข ตั้งแต่ปี 2015 จากผลการทดลองโดยให้ค่า Accuracy 99.5%

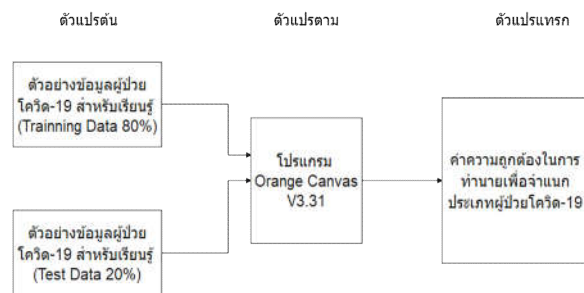
COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm [22] งานวิจัยนี้ได้นำเสนอโมเดล Random Forest ที่ปรับแต่งอย่างละเอียด โดยใช้ข้อมูลทางภูมิศาสตร์ การเดินทาง สุขภาพ และข้อมูลประชากรของผู้ป่วยโควิด-19 เพื่อคาดการณ์ความรุนแรงของเคสและผลลัพธ์ที่เป็นไปได้ การฟื้นตัว หรือการเสียชีวิต โดยข้อมูลที่ใช้จะเป็นข้อมูลแบบเปิดชื่อ “Novel Corona Virus 2019 Dataset” จากเว็บไซต์ www.kaggle.com โดยโมเดล Random Forest ได้ให้ค่า Accuracy อยู่ที่ 94%

Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results [23] ได้ทำการวิจัยและศึกษาจากผลการตรวจเลือดเพื่อประเมินผู้ต้องสงสัยในการติดเชื้อโรคโควิด-19 ในการศึกษาได้รวบรวมตัวอย่างทั้งหมด 253 ตัวอย่าง จากผู้ป่วยต้องสงสัย 169 รายจากในหลายๆ แห่ง รวมทั้งในโรงพยาบาลปอดหกลานใจที่ประเทศจีน โดยทำการตรวจเลือด 49 ราย โดยจะนำค่าจากผลตรวจเลือด 11 ค่ามาวิเคราะห์ผ่านโปรแกรม Covid-19 Assistant Discrimination 2.0 เพื่อทำนายโอกาสการติดโรคโควิด-19 จากการศึกษาได้ใช้โมเดล Random forest โดยให้ค่า Accuracy 95.95% และ Specificity อยู่ที่ 96.95%

อย่างไรก็ตาม งานวิจัยที่ได้กล่าวมาไม่ได้มีวัตถุประสงค์มุ่งเน้นการคัดกรองผู้ป่วยที่มีความเสี่ยงในการติดโรคโควิด-19 ภายใต้ระยะเวลาอันรวดเร็ว งานวิจัยชิ้นนี้จึงนำเสนอการประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่องเพื่อแยกประเภทของผู้ป่วยและผู้ป่วยในกลุ่มเสี่ยงของโรคโควิด-19 โดยผู้วิจัยได้ทำการสรุปงานวิจัยที่เกี่ยวข้องและแยกเป็น 2 หมวดหมู่ด้วยกันคือ 1. จำแนกผู้ป่วยโรคโควิด-19 จากภาพเอกซเรย์ทรวงอก 2. จำแนกผู้ป่วยโรคโควิด-19 จากชุดข้อมูลหลากหลายฟีเจอร์ (Feature) โดยทุกงานจะมีการใช้อัลกอริทึม (Algorithm) ที่แตกต่างกันออกไป เป็นงานวิจัยเพื่อวัตถุประสงค์ในการคัดแยก - จำแนกผู้ป่วย , ทำนายโอกาสการติดเชื้อโควิดหรือทำนายผู้ต้องสงสัยว่าติดเชื้อหรือไม่ ซึ่งงานวิจัยฉบับนี้ที่ผู้วิจัยได้ให้ความสำคัญกับหลังเหตุการณ์เมื่อพบผู้ติดเชื้อโควิด-19 ในการคัดแยกผู้ป่วยโควิด-19 เพื่อช่วยคัดแยกและกรองผู้ต้องสงสัยได้อย่างรวดเร็ว ซึ่งจะช่วยเหลือเวลาได้เร็วขึ้น [6]

3. วิธีดำเนินการวิจัย

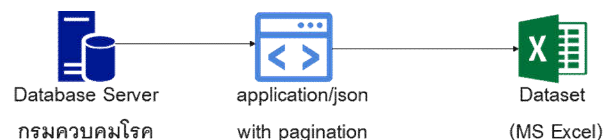
แนวคิดหลักการวิจัยนี้คือการนำเทคนิคการเรียนรู้ของเครื่องมาประยุกต์ใช้ในการคัดแยกผู้ป่วยโควิด-19 โดยอัลกอริทึมที่ใช้ในการพัฒนาแบบจำลองจะประกอบไปด้วย 3 เทคนิคคือ Neural Network , Naive bayes และ Random Forest และหาค่าความถูกต้องหรือค่าความแม่นยำมากที่สุดจาก 3 เทคนิค



รูปที่ 1. กรอบแนวคิด

3.1 Data Set (วิธีการรวบรวมข้อมูล)

ทำการรวบรวมข้อมูลจากศูนย์กลางข้อมูลภาครัฐ ซึ่งเป็นข้อมูลแบบเปิดจากกรมควบคุมโรค กระทรวงสาธารณสุข โดยจะเป็นข้อมูลผู้ป่วยโควิด-19 ระลอกที่ 1 ถึงระลอกที่ [7] ตั้งแต่ 1 มกราคม 2563 ถึง วันที่ 1 ตุลาคม 2564 เป็นจำนวน 1,608,923 ราย



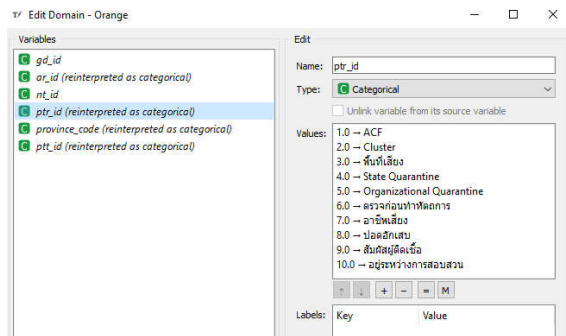
รูปที่ 2. ขั้นตอนการพัฒนาส่วนการดาวน์โหลดข้อมูล

3.2 Data Preparation (การเตรียมข้อมูล)

Formatting (การจัดรูปแบบข้อมูล) ขั้นตอนนี้เป็นขั้นตอนที่ทำการแปลงข้อมูลที่ได้ทำการเก็บรวบรวมมา ให้กลายเป็นข้อมูลที่สามารถนำไปวิเคราะห์ในขั้นถัดไปได้ ในการแปลงข้อมูลนี้อาจจะต้องมีการทำข้อมูลให้ถูกต้อง เช่นการแปลงข้อมูลให้อยู่ในช่วง (scale) เดียวกัน โดยจะทำการแปลงข้อมูลจากตัวอักษร (text) เป็น ตัวเลข (numeric) เช่น Attributes และ Values ทั้งหมด รวมทั้งจัดกลุ่ม Attributes ของชุดข้อมูล เช่น gender, age_range, nationality, risk, patient_type และ province

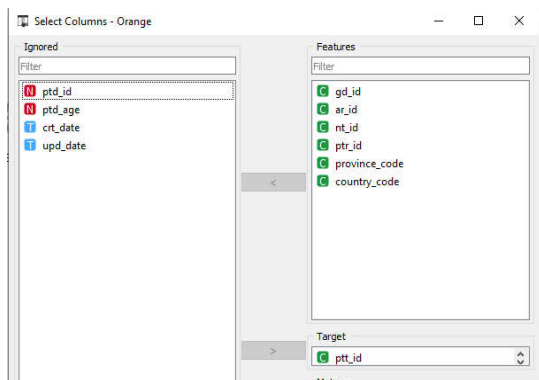
Data Cleaning (การทำความสะอาดข้อมูล) เนื่องจากคะแนนหรือข้อมูลบางส่วนไม่สามารถใช้ในการคำนวณได้ เพราะในการแบ่งข้อมูลถูกแบ่งออกเป็นสองส่วนคือข้อมูลที่มี

การซ้อนทับและข้อมูลที่ไม่มีการซ้อนทับนั้นมีการหาระยะห่างระหว่างข้อมูล และข้อมูลที่ใช้จะต้องเป็นข้อมูลที่ไม่มีการสูญหายหรือ Missing Value เพราะข้อมูลดังกล่าวจะไม่สามารถหาระยะห่างระหว่างข้อมูลได้



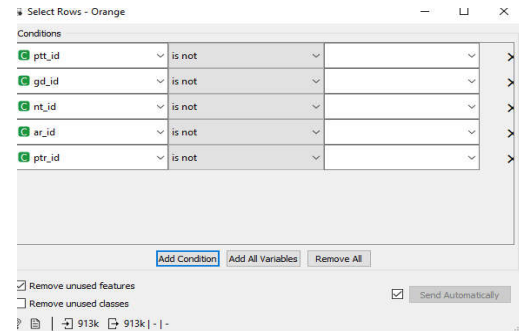
รูปที่ 3. การกำหนด Label และประเภทข้อมูล

การกำหนดประเภทข้อมูล age_range , nationality , risk , province และ patient_type ใน Edit Domain Widget ให้เป็นประเภท Categorical เนื่องจากเป็นข้อมูลที่มีการถูกจัดกลุ่ม จากนั้นจะทำการกำหนด Label ให้กับข้อมูลต่างๆ เพื่อการแสดงผลที่เหมาะสมในการวิเคราะห์ข้อมูลเช่น 1 = ACF , 2 = Cluster เป็นต้น ดังรูปที่ 3



รูปที่ 4. การเลือก Target , Features

ทำการเลือก age_range , nationality , risk , province และ patient_type ที่เป็นข้อมูลประเภท Categorical เท่านั้น โดยจะตัดส่วน txn_date และ update_date ออกจากการวิเคราะห์เนื่องจากเป็นตัวแปรต่อเนื่องไม่สามารถแยกจากกันได้ ดังรูปที่ 4



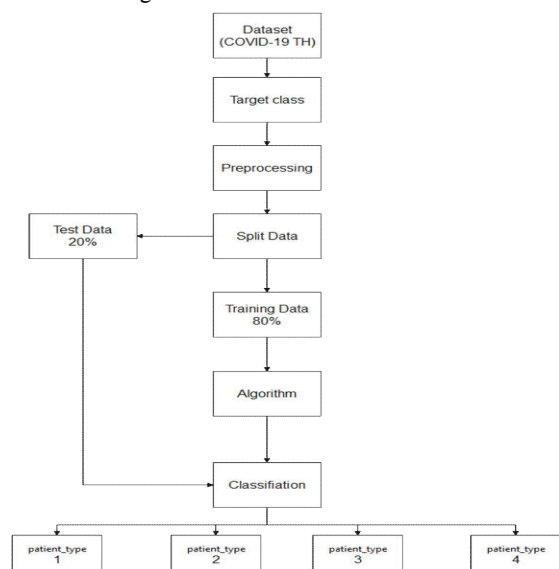
รูปที่ 5. การกำจัด Missing value

ทำการกำหนดเงื่อนไข (Condition) ในการคัดกรองข้อมูลที่เป็นค่าว่างหรือข้อมูลที่ไม่สามารถวิเคราะห์ได้ ดังรูปที่ 5

3.3 Modeling (การสร้างโมเดล)

กระบวนการนี้ผู้วิจัยได้ใช้โปรแกรม Orange Canvas ในการสร้างตัวแบบข้อมูลสำหรับการเรียนรู้และทดสอบเพื่อเปรียบเทียบโมเดลที่ดีที่สุด ก่อนจะนำโมเดลนั้นไปพัฒนาเว็บแอปพลิเคชันในการคัดแยกผู้ป่วยโควิด-19

จากการตั้งค่าพารามิเตอร์ที่ดีที่สุดสำหรับข้อมูลจากกรมควบคุมโรค กระทรวงสาธารณสุข โดยจะเป็นข้อมูลผู้ป่วยโควิด-19 ระลอกที่ 1 ถึงระลอกที่ 3 และทำงานร่วมกับโมเดลด้วยวิธีการรวมผลลัพธ์ที่ได้ค่าความถูกต้องมากที่สุด ซึ่งอยู่ในรูปแบบของเทคนิคโมเดล Random Forest , Neural network และ Naive Bayes โดยได้ใช้วิธี Split test ในการแบ่งข้อมูลสำหรับเป็นชุดการเรียนรู้ 80% และชุดทดสอบ 20% แล้วนำเข้าโปรแกรม Orange Canvas



รูปที่ 6. โมเดลการวิเคราะห์คัดแยกผู้ป่วยโควิด-19

รูปที่ 6 ขั้นตอนนี้จะทำการสร้างโมเดลด้วยวิธี Naive Bayes , Random Forest และ Neural network โดยแบ่งข้อมูลจำนวน 80% เป็นข้อมูลสำหรับสร้างโมเดลและข้อมูลจำนวน 20% ใช้เป็นข้อมูลสำหรับทดสอบโมเดล ในงานวิจัยนี้ได้สร้างโมเดลแยกตามประเภทผู้ป่วยโดย Class ของข้อมูลสามารถดูได้ตามตารางที่ 2

3.4 Evaluation (การวัดประสิทธิภาพโมเดล)

หลังจากที่แบ่งข้อมูลออกเป็น 2 ส่วนและใช้โมเดลที่สร้างได้จากข้อมูลจำนวน 80% มาทำการคัดแยกผู้ป่วยโควิด-19 กับข้อมูลผู้ป่วยที่อยู่ใน 20% ที่เหลือ และใช้วัดค่าความถูกต้องโดยใช้ Confusion matrix ตามหัวข้อที่ 2.7 เพื่อดูว่าเทคนิควิธี Naive Bayes , Random Forest และ Neural network เทคนิคที่สร้างโมเดลไหนให้ค่าผลทำนายดีมากที่สุดจึงจะนำโมเดลนั้นไปใช้งานต่อไป ค่าความแม่นยำในการจำแนกประเภทข้อมูลของ Positive นั้นสามารถคำนวณได้จากสมการที่ 1

$$\text{True Positive Rate (TPR)} = \frac{TP}{TP + FN} \quad (1)$$

ค่าความแม่นยำในการจำแนกประเภทข้อมูลของ Negative นั้นสามารถคำนวณได้จากสมการที่ 2

$$\text{True Negative Rate (TNR)} = \frac{TN}{FP + TN} \quad (2)$$

ค่า Precision เป็นการวัดความแม่นยำของโมเดล โดยพิจารณาแยกทีละ Class ดังสมการที่ 3

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

ค่า Recall เป็นการวัดความถูกต้องของโมเดล โดยพิจารณาแยกทีละ Class ดังสมการที่ 4

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

ค่า F-measure เป็นการวัดค่า Precision และ Recall พร้อมกันของโมเดลโดยพิจารณาแยกทีละ Class ดังสมการที่ 5

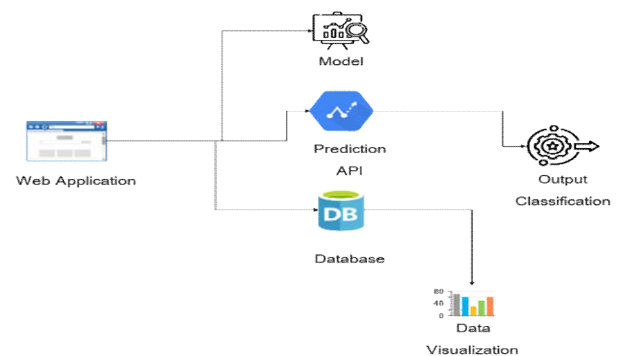
$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

ค่า Accuray เป็นการวัดความถูกต้องของโมเดลโดยพิจารณา รวมทุก Class คือ จำนวน True Positive (TP) ของทุก Class รวมกัน ดังสมการที่ 6

$$\text{Accuray} = \frac{TP + TN}{TP + FN + FP + TN} \quad (6)$$

3.5 Deployment (การนำโมเดลไปใช้งาน)

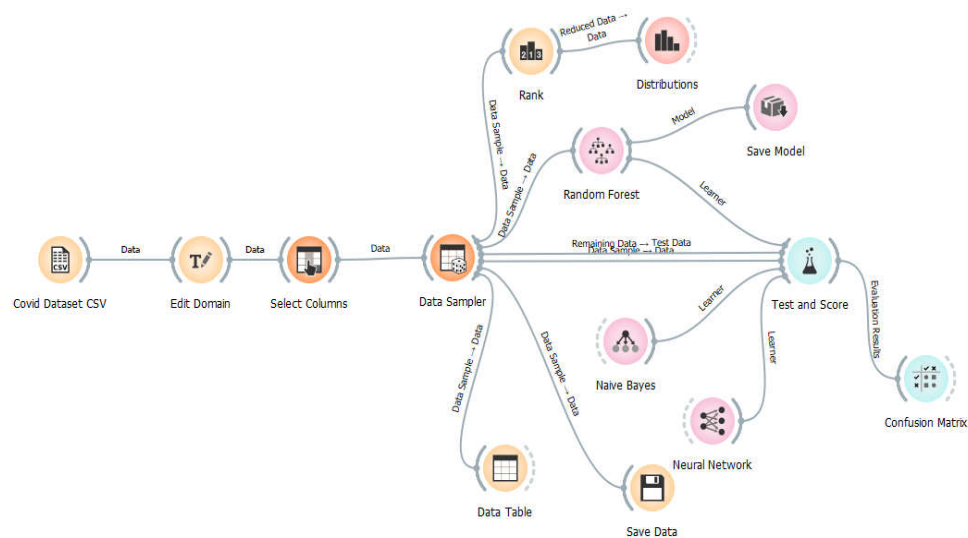
จากโมเดลที่สร้างโดย Naive Bayes , Random Forest และ Neural network ที่สร้างขึ้นมาได้นั้นสามารถนำไปทำการคัดแยกผู้ป่วยโควิด-19 หลังจากนั้นจะต้องพัฒนา Prediction API ของโมเดลที่สร้างขึ้นมาเพื่อเป็นเว็บเซอร์วิส ในการรองรับการส่งข้อมูลจากเว็บแอปพลิเคชันในการคัดแยกผู้ป่วยโควิด-19 ดังรูปที่ 7



รูปที่ 7. Prediction API ในการคัดแยกผู้ป่วยโควิด-19

4. ผลการวิเคราะห์ข้อมูล

งานวิจัยนี้มีข้อมูลจำนวน 1,608,923 เรคคอร์ด หลังจากผ่านกระบวนการ Data Preparation แล้วข้อมูลจะอยู่ที่จำนวน 913,152 เรคคอร์ด ประกอบด้วย คุณลักษณะดังนี้ 1) คำนำนักคุณลักษณะของข้อมูลจำนวน 5 คุณลักษณะ คือ gender , age_rang, nationality , province , risk และ 2) คลาส (class) ที่ใช้ในการวิจัยประกอบไปด้วย 4 คลาส มีจำนวนข้อมูลในแต่ละคลาสดังนี้ คลาสที่ 1 มีจำนวน 587,079 เรคคอร์ด คลาสที่ 2 มีจำนวน 2,755 เรคคอร์ด คลาสที่ 3 มีจำนวน 52,548 เรคคอร์ด และคลาสที่ 4 มีจำนวน 270,770 ดังตารางที่ 4



รูปที่ 8. การใช้โปรแกรม Orange Canvas เพื่อเปรียบเทียบ โมเดลการเรียนรู้ของเครื่อง

ตารางที่ 4. อัตราส่วนของข้อมูลในแต่ละ Class

Class	จำนวน	อัตราส่วน
1	587,049	64.29%
2	2,755	0.30%
3	52,547	5.75%
4	270,770	29.65%

ตารางที่ 5. Information Gain ของฟีเจอร์ในการวิจัย

ตัวแปรที่	Feature	Information Gain
1	gender	0.003
2	age_rang	0.001
3	nationality	0.019
4	province	0.213
5	risk	0.816

จากตารางที่ 5 พบว่าการคัดเลือกฟีเจอร์ทั้งหมด โดยใช้เทคนิค คำนวณค่าน้ำหนักของฟีเจอร์ต่างๆ โดยฟีเจอร์ที่มีค่าน้ำหนักมากที่สุดคือ risk ค่าน้ำหนักที่ 0.816 และ ฟีเจอร์ที่มีค่าน้ำหนักน้อยที่สุดคือ gender ค่าน้ำหนักที่ 0.003

4.1 ผลการจำแนกข้อมูลด้วย Random Forest

ตารางที่ 6. ตัววัดประสิทธิภาพของอัลกอริทึม Random Forest

Performance Classification (Random Forest)					
	true(1)	true(2)	true(3)	true(4)	Precision
pred.(1)	459329	19	5254	5230	97.76%
pred.(2)	6	2176	7	27	98.19%
pred.(3)	4091	16	28188	9725	67.08%
pred.(4)	20976	29	3524	191925	88.66%
Recall	94.82%	97.14%	76.23%	92.75%	

จากตารางที่ 6 พบว่า Class 1 True Positive เท่ากับ 459,329 ค่า False Positive เท่ากับ 10,053 ค่า คิดเป็น Precision เท่ากับ 97.76% และ Recall เท่ากับ 94.82% Class 2 True Positive เท่ากับ 2,176 ค่า False Positive เท่ากับ 40 ค่า คิดเป็น Precision เท่ากับ 98.19% และ Recall เท่ากับ 94.14% Class 3 True Positive เท่ากับ 28,188 ค่า False Positive เท่ากับ 13,832 ค่า คิดเป็น Precision เท่ากับ 67.08% และ Recall เท่ากับ 76.23% Class 4 True Positive เท่ากับ 191,925 ค่า False Positive เท่ากับ 24,529 ค่า คิดเป็น Precision เท่ากับ 88.66 % และ Recall เท่ากับ 92.75% สรุปเป็นค่า Accuracy ได้ 93.33%

4.2 ผลการจำแนกข้อมูลด้วย Neural Network

ตารางที่ 7. ตัววัดประสิทธิภาพของอัลกอริทึม Neural Network

Performance Classification (Neural Network)					
	true(1)	true(2)	true(3)	true(4)	Precision
pred.(1)	455433	19	4544	9836	96.93%
pred.(2)	12	2198	0	6	99.18%
pred.(3)	3408	14	24653	13945	58.66%
pred.(4)	19387	39	1932	195096	90.13%
Recall	95.23%	96.82%	79.19%	89.13%	

จากตารางที่ 7 พบว่า Class 1 True Positive เท่ากับ 455,433 ค่า False Positive เท่ากับ 14,399 ค่า คิดเป็น Precision เท่ากับ 96.93% และ Recall เท่ากับ 95.23% Class 2 True Positive เท่ากับ 2,198 ค่า False Positive เท่ากับ 18 ค่า คิดเป็น Precision เท่ากับ 99.18% และ Recall เท่ากับ 96.82% Class 3 True Positive เท่ากับ 24,653 ค่า False Positive เท่ากับ 17,367 ค่า คิดเป็น Precision เท่ากับ 58.66% และ Recall เท่ากับ 79.19% Class 4 True Positive เท่ากับ 195,096 ค่า False Positive เท่ากับ 21,449 ค่า คิดเป็น Precision เท่ากับ 90.13% และ Recall เท่ากับ 89.13% สรุปเป็นค่า Accuracy ได้ 92.7%

4.3 ผลการจำแนกข้อมูลด้วย Naive Bayes

ตารางที่ 8. ตัววัดประสิทธิภาพของอัลกอริทึม Naive Bayes

Performance Classification (Naive Bayes)					
	true(1)	true(2)	true(3)	true(4)	Precision
pred.(1)	455784	14	7203	6831	97%
pred.(2)	28	2113	3	72	95.35%
pred.(3)	4147	7	24724	13142	58.83%
pred.(4)	21009	36	5060	190349	87.93%
Recall	94.76%	97.14%	66.83%	90.47%	

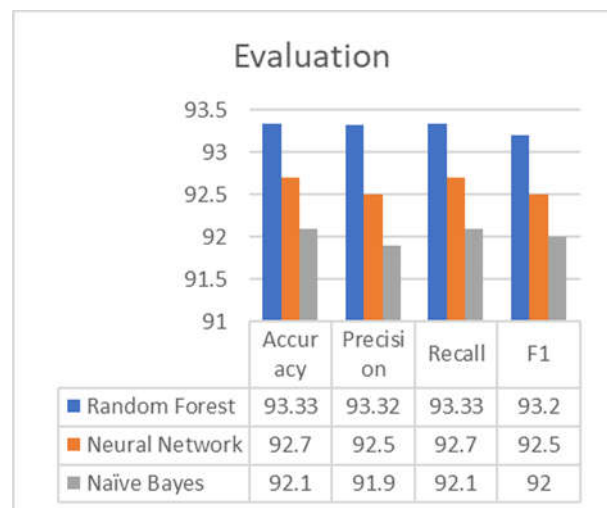
จากตารางที่ 8 พบว่า Class 1 True Positive เท่ากับ 455,784 ค่า False Positive เท่ากับ 14,048 ค่า คิดเป็น Precision เท่ากับ 97% และ Recall เท่ากับ 94.76% Class 2 True Positive เท่ากับ 2,113 ค่า False Positive เท่ากับ 103 ค่า คิดเป็น Precision เท่ากับ

95.35% และ Recall เท่ากับ 97.14% Class 3 True Positive เท่ากับ 24,724 ค่า False Positive เท่ากับ 17,296 ค่า คิดเป็น Precision เท่ากับ 58.83% และ Recall เท่ากับ 66.83% Class 4 True Positive เท่ากับ 190,349 ค่า False Positive เท่ากับ 26,105 ค่า คิดเป็น Precision เท่ากับ 87.93% และ Recall เท่ากับ 90.47% สรุปเป็นค่า Accuracy ได้ 92.1%

4.4 สรุปผลการเปรียบเทียบค่าความแม่นยำในแต่ละเทคนิค

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อทดลองสร้างโมเดลการทำนายโดยใช้ข้อมูลผู้ป่วยโควิด-19 จากกรมควบคุมโรค ร่วมกับเทคนิคการเรียนรู้ของเครื่องได้แก่ Random Forest , Neural network และ Naive Bayes เพื่อใช้ในการคัดแยกผู้ป่วยโควิด-19 ในการเปรียบเทียบประสิทธิภาพความแม่นยำในการทำนาย โดย Random Forest มีค่า Accuracy 93.33% ส่วน Neural Network มีค่า Accuracy 92.7% และ Naive Bayes มีค่า Accuracy ที่ 92.1%

ผู้วิจัยจึงใช้เทคนิค Random Forest เนื่องจากให้ค่าความถูกต้องและแม่นยำมากที่สุดมาสร้างโมเดลในการทำนายเพื่อคัดแยกผู้ป่วยโควิด-19 ต่อไป ดังรูปที่ 9

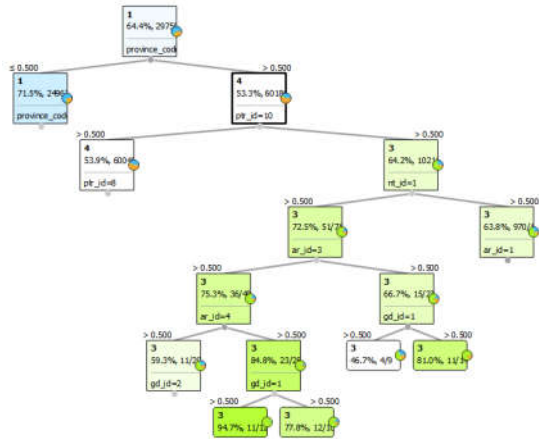


รูปที่ 9. การเปรียบเทียบความแม่นยำในแต่ละอัลกอริทึม

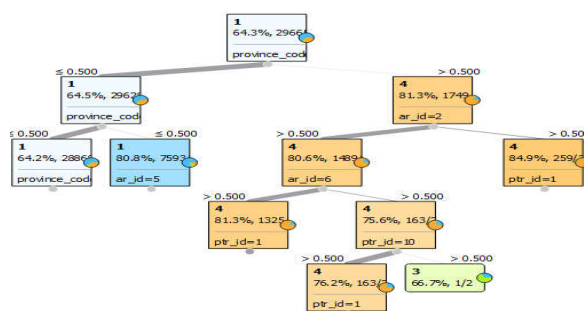
4.5 สรุปผลการทำงานของ Random Forest

โมเดล Random Forest Classification ให้ผลการทำนายที่มีความแม่นยำมากที่สุดถึง 93.33 % แต่ก็ขึ้นอยู่กับข้อมูลความเสี่ยงและจังหวัดในการวัดผลด้วย เนื่องจากในแต่ละจังหวัดยังมีความเสี่ยงที่ใกล้เคียงกัน โดยเฉพาะจังหวัดกรุงเทพมหานครและ

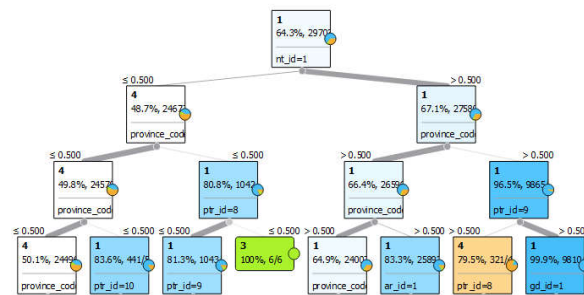
ปริมาณผล แต่ในต่างจังหวัดก็จะมีความเสี่ยงที่แตกต่างกัน
ออกไป ซึ่งจะเห็นได้ว่าการใช้งาน Random Forest
Classification สามารถใช้งานได้หลากหลายรูปแบบ



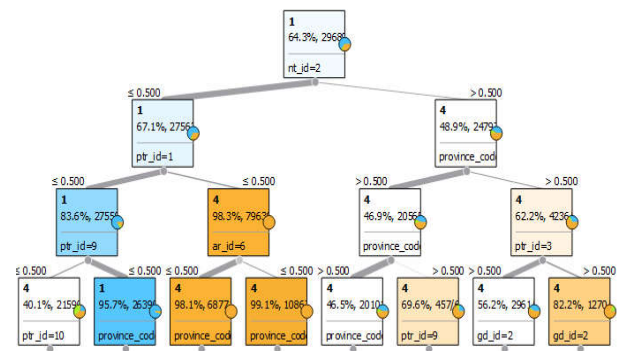
รูปที่ 10. Decision Tree ที่ 1



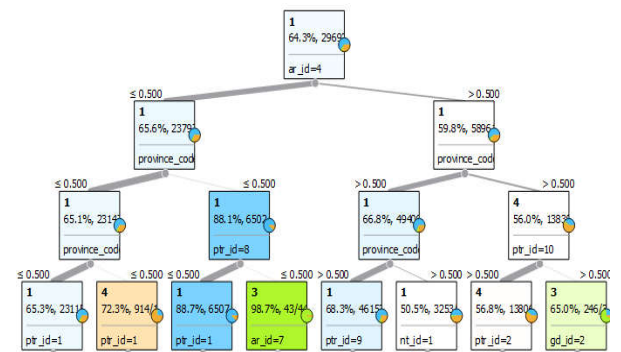
รูปที่ 11. Decision Tree ที่ 2



รูปที่ 12. Decision Tree ที่ 3



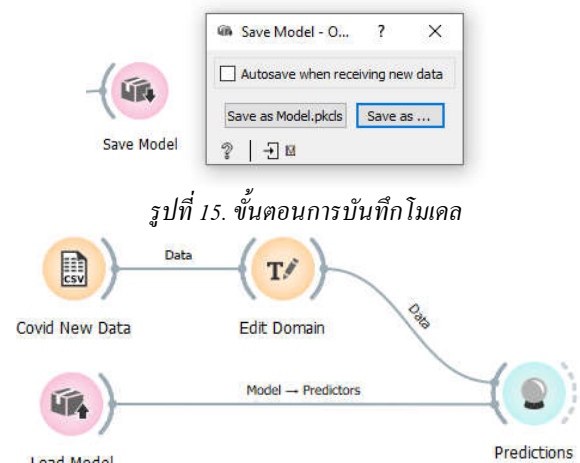
รูปที่ 13. Decision Tree ที่ 4



รูปที่ 14. Decision Tree ที่ 5

4.6 การนำโมเดลไปใช้งาน (Deployment)

จากรูปที่ 8 เมื่อได้ผลจากการหาประสิทธิภาพของเทคนิคทั้ง 3
เทคนิคแล้ว นั่นคือได้ Random Forest จึงมีกระบวนการบันทึก
โมเดลโดยจะได้ไฟล์ Model.pkcls เพื่อนำไปใช้งาน ดังรูปที่ 15



รูปที่ 15. ขั้นตอนการบันทึกโมเดล

เมื่อมีชุดข้อมูลใหม่สามารถใช้เครื่องมือ Load Model Widget
ในโปรแกรม Orange Canvas ในการทำนายผลได้ทันทีดังรูปที่

```
1 import pickle
2 # ทำการ Load โมเดลที่เก็บไว้
3 model = pickle.load(open("Model.pkcls", "rb"))
4 # ข้อมูลผู้ป่วยใหม่
5 newdata = [[1, 8, 1, 9, 10]]
6 # ทำนายการคัดแยกผู้ป่วย
7 predict = model.predict(newdata)
8 print(model)
```

รูปที่ 17. การนำโมเดลไปใช้งานด้วย Python

รูปที่ 17 ด้วยผลการทดสอบทั้งหมดจะได้โมเดลด้วยไฟล์ Model.pkcls สำหรับการนำไปพัฒนาระบบคัดแยกผู้ป่วยโควิด 19 ด้วยภาษา Python ต่อไป

4.7 ระบบคัดแยกผู้ป่วยโควิด-19

กระบวนการพัฒนาระบบได้ใช้ FastAPI Framework ด้วยภาษา Python เพื่อสร้าง Web Service API ในการทำนายข้อมูลโดยได้นำโมเดลที่ได้ไป Deployment เข้ากับระบบ และใช้ฐานข้อมูล Postgresql จากนั้นทำการพัฒนาเว็บแอปพลิเคชันในการคัดแยกผู้ป่วยโควิด-19 ดังรูปที่ 18-23

แบบฟอร์มวิเคราะห์ข้อมูลผู้ป่วย

วันที่เริ่มป่วย: 04/01/2022 เพศ: ชาย อายุ: 28

สัญชาติ: ไทย ความเสี่ยง: พื้นที่เสี่ยง

จังหวัด: กรุงเทพมหานคร

ทำนายผล

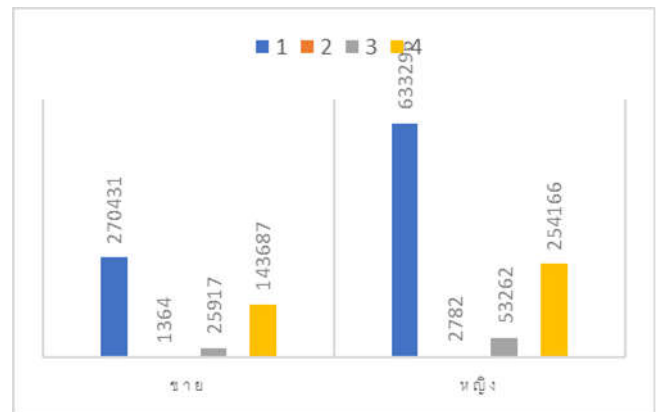
รูปที่ 18. หน้าจอทำนายการคัดแยกผู้ป่วยโควิด-19



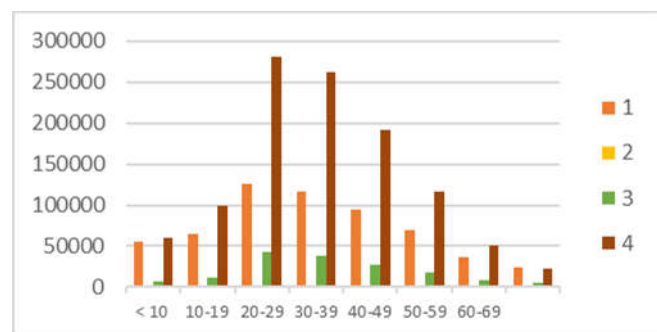
รูปที่ 19. ผลการทำนายในการคัดแยก



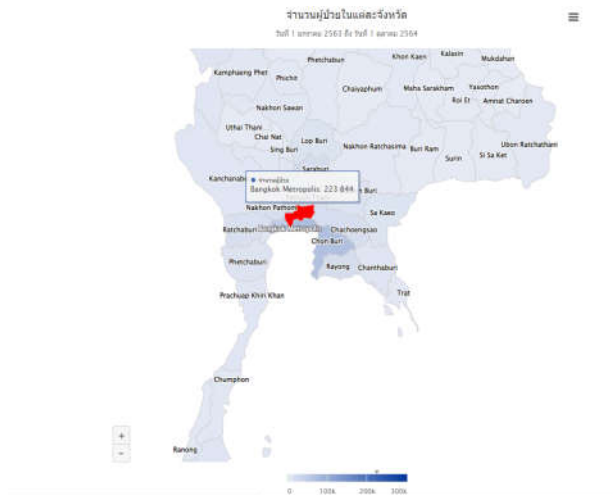
รูปที่ 20. กราฟแสดงยอดผู้ป่วยในแต่ละเดือน



รูปที่ 21. อัตราส่วนเพศต่อประเภทผู้ป่วย



รูปที่ 22. อัตราส่วนช่วงอายุต่อประเภทผู้ป่วย



รูปที่ 23. จำนวนผู้ป่วยในแต่ละจังหวัด

5. บทสรุปและการอภิปราย

การเปรียบเทียบประสิทธิภาพของอัลกอริทึมและการคัดเลือกคุณลักษณะที่เหมาะสมเพื่อการทำนายในการคัดแยกผู้ป่วยโควิด-19 โดยในการดำเนินงานเริ่มต้นได้ทำการศึกษาอัลกอริทึมต่างๆ ของการเรียนรู้ของเครื่อง การรวมรวมข้อมูล วิธีในการ Training Data รูปแบบ Test Data ซึ่งการวิเคราะห์สังเคราะห์จากข้อมูล และเอกสารกำหนดขอบเขต ทำการออกแบบแผนการทดลองโดยใช้โปรแกรม Orange Canvas ในการสร้างโมเดลตามเทคนิคต่างๆ ในการทดลอง และประมวลผลผลลัพธ์ ทำให้ระยะเวลาในการทดลองมีความเร็ว และมีประสิทธิภาพมากยิ่งขึ้น หลังจากรวบรวมข้อมูลและผ่านการค้นคว้าและปรับคุณลักษณะของข้อมูลแล้วก็นำข้อมูลไปทดลองในการทำนายการคัดแยกผู้ป่วยโควิด-19 ได้ทั้ง 3 เทคนิค Naive Bayes , Random Forest และ Neural network ซึ่งทำการเปรียบเทียบเพื่อให้ได้วิธีการที่ได้ค่าความแม่นยำและถูกต้องมากที่สุด จากนั้นหาตัวแบบเทคนิคในการทำนายที่ให้ค่าความถูกต้องมากที่สุดที่เหมาะสมกับการนำไปใช้ในการทำนายการคัดแยกผู้ป่วยโควิด-19 ผลการทดลองในการวิจัยพบว่า Random Forest ให้ค่าความถูกต้องสูงที่สุดที่ร้อยละ 93.30% Neural network ให้ค่าความถูกต้องสูงที่สุดที่ร้อยละ 92.7% และ Naive Bayes ให้ค่าความถูกต้องสูงที่สุดที่ร้อยละ 92.1% ตามลำดับ สามารถบอกได้ว่าผู้ป่วยโควิด-19 มีโอกาสถูกจัดให้อยู่ประเภทใด และสามารถนำไปเขียนโปรแกรมเพื่อพัฒนาระบบให้มีการคัดแยกผู้ป่วยโควิด-19 ได้อย่างถูกต้องมากยิ่งขึ้น และลดข้อผิดพลาดของข้อมูลให้น้อยลง

การทำนายด้วย 3 เทคนิคที่เลือกใช้ในงานวิจัยนี้ นอกจากจะสามารถทำการ Classification ได้ดีแล้ว และยังสามารถทำได้ Regression ได้ดีเช่นกัน

6. ข้อเสนอแนะ

ข้อมูลที่ผู้วิจัยรวบรวมจากกรมควบคุมโรคนั้นพบว่าข้อมูลในแต่ละคุณลักษณะยังมีส่วนที่ขาดและไม่ครบข้างในบางราย คือคุณลักษณะ อาชีพของผู้ป่วย ซึ่งเป็นข้อมูลที่ขาดไปมากคิดเป็นร้อยละ 1.72% จากข้อมูลที่รวบรวมได้ทั้งหมด จึงไม่สามารถนำไปวิเคราะห์ได้ อีกทั้งข้อมูลประเทศยังมีการรวบรวมในลักษณะซ้ำซ้อน การใช้คำที่มีความหมายเหมือนกัน แต่ข้อความต่างกันเล็กน้อย จึงทำให้ข้อมูลไม่เป็นระเบียบ ในการวิจัยครั้งต่อไปควรเพิ่มคุณลักษณะของข้อมูลและทดลองอัลกอริทึมอื่นๆ เพิ่มเติม

เอกสารอ้างอิง

- [1] S. Manmana, S. Iamsirithaworn, and S. Uttayamakul, "Coronavirus Disease-19 (Covid-19)," Journal of Bamrasnaradura Infectious Diseases Institute," vol. 14(2), pp. 124–33, Mar. 2020.
- [2] T. Saowapa, P. Supistra, and R. Ranistha, "Nursing care for patients with COVID-19 in the isolation unit, Siriraj Hospital." Jul. 2020. [Online]. Available: <https://dx.doi.org/10.33192/Simedbull.2020.29>
- [3] T.-N. Rachada, "COVID-19: An Invisible War Against Coronavirus," THAI FOOD AND DRUG JOURNAL, vol. 27(2), May 2020.
- [4] K. Natawan, "Knowledge, Attitudes, and Preventive Behaviors of COVID-19 among People Living in Amphoe U- thong, Suphanburi Province," Journal of

- Prachomklao College of Nursing, Phetchaburi Province, vol. 4 No.1, Apr. 2021.
- [5] K. L. Trepanning and J. Nammong, “Factors Associated with Preventive Behaviors towards the Coronavirus 2019 (COVID-19) of Employees in a Large Factory Krathumbean District, Samutsakorn Province,” Journal of Nursing, Siam University, 2021.
- [6] Kasetsart university, “Covid-19 and epidemiology,” COVID-19 Course, Jul. 2020. <https://learningcovid.ku.ac.th/course/?c=3&l=1>
- [7] Department of disease control, “DDC OPENDATA Covid19 Thailand,” DDC OPENDATA Covid19 Thailand, 2021. <https://covid19.ddc.moph.go.th/> (accessed Oct. 22, 2021).
- [8] A. Puitrakul, “What is machine learning,” What is machine learning, Jan, 2018. <https://arnondora.in.th/what-is-machine-learning/>
- [9] E. Digest, “Artificial intelligence (AI) -- a catalyst for China’s post-COVID economy,” Kasikorn research center, Jun, 2020. <https://www.kasikornresearch.com/th/analysis/k-social-media/Pages/AI-China-FB300620.aspx> (accessed Oct. 25, 2021).
- [10] S. Lalmuanawma, J. Hussain, and L. Chhakchhuak, “Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review,” Chaos, Solitons & Fractals, vol. 139, p. 110059, Oct. 2020, doi: 10.1016/j.chaos.2020.110059.
- [11] Department of disease control, “Coronavirus Disease 2019: COVID-19.” Dec. 2563. [Online]. Available: https://ddc.moph.go.th/viralpneumonia/file/g_srrt/g_srrt_241263.pdf
- [12] Department of disease control, “Guideline for disease control in quarantine facilities.” Containment Measures Mission Group (QUARANTINE) Emergency Operations Center ,Department of disease control, Jul. 2020. [Online]. Available: https://ddc.moph.go.th/viralpneumonia/file/g_quarantine/g_quarantine_state210763n.pdf
- [13] C. Patcharachoenwong, H. Kankawee, and K. Warangkhan, “Arrival Time Prediction Model to a Pier for Public Transportation Boats,” Journal of Science Ladkrabang, vol. 29(2), 2020, [Online]. Available: https://li01.tci-thaijo.org/index.php/science_kmitl/article/download/241105/169801/
- [14] E. Pacharawongsakda, An introduction to Data Mining Techniques, vol. 2. 2014.
- [15] A. Thongjit, P. Suksawang, and M. Jatupat, “Development of Data Classification using a Hybrid Method of Adaptive Artificial Neural Networks and Particle Swarm Optimization for Identifying Patients at Risk of Diabetes,” Research Methodology & Cognitive Science, vol. 17(2), Dec. 2019, [Online]. Available: <https://so05.tci-thaijo.org/index.php/RMCS/article/download/243179/165072/>
- [16] P. Duangklang and R. Kruakaew, “Models for automatic aircraft type prediction,” NKRAFA Journal of Science and Technology, 2019.

- [17] “Confusion Matrix - an overview | ScienceDirect Topics.”
<https://www.sciencedirect.com/topics/engineering/confusion-matrix> (accessed Jun. 08, 2022).
- [18] A. A. Ardakani, A. R. Kanafi, U. R. Acharya, N. Khadem, and A. Mohammadi, “Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks,” *Computers in Biology and Medicine*, vol. 121, p. 103795, Jun. 2020, doi: 10.1016/j.combiomed.2020.103795.
- [19] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. Rajendra Acharya, “Automated detection of COVID-19 cases using deep neural networks with X-ray images,” *Computers in Biology and Medicine*, vol. 121, p. 103792, Jun. 2020, doi: 10.1016/j.combiomed.2020.103792.
- [20] L. Sun et al., “Combination of four clinical indicators predicts the severe/critical symptom of patients infected COVID-19,” *Journal of Clinical Virology*, vol. 128, p. 104431, Jul. 2020, doi: 10.1016/j.jcv.2020.104431.
- [21] Z. Karhan and F. Akal, “Covid-19 Classification Using Deep Learning in Chest X-Ray Images,” in *2020 Medical Technologies Congress (TIPTEKNO)*, Antalya, Turkey, Nov. 2020, pp. 1–4. doi: 10.1109/TIPTEKNO50054.2020.9299315.
- [22] C. Iwendi et al., “COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm,” *Frontiers in Public Health*, vol. 8, p. 357, 2020, doi: 10.3389/fpubh.2020.00357.
- [23] J. Wu et al., “Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results,” Apr. 2020. doi: 10.1101/2020.04.02.20051136.