การเปรียบเทียบประสิทธิภาพตัวแบบพยากรณ์ความเสี่ยงการติดเชื้อโควิด-19 ด้วยเทคนิคเหมืองข้อมูล

จิรายุ สิทธิชัย[1*] และ ภรัณยา ปาลวิสุทธิ์ [1]

[1]สาขาวิชาวิทยาการข้อมูล คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม
[*]Jirayu7991@gmail.com

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองในการทำนายหาโอกาสที่จะติดเชื้อโควิด19 โดยใช้เทคนิคเหมืองข้อมูล 3 วิธี คือ วิธีต้นไม้ตัดสินใจ วิธีแบบเบย์และวิธีการเพื่อนบ้านใกล้ที่สุด นำมาเปรียบเทียบประสิทธิภาพของโมเดลการจำแนกหาตัวแบบที่เหมาะสมเพื่อใช้ทำนายหาโอกาสที่จะติดเชื้อโควิด-19 โดยใช้ข้อมูล COVID-19 จำนวน 5,434 แถว 21 คอลัมน์ จากเว็บไซต์ kaggle วิเคราะห์ข้อมูลบนพื้นฐานของวิธี CRISP-DM โดยใช้โปรแกรม RapidMiner ในการสร้างแบบจำลอง ในการวิเคราะห์ข้อมูลจะหาค่าความถูกต้อง ความแม่นยำ และความระลึก

ผลการศึกษาพบว่าวิธีต้นไม้ตัดสินใจ มีประสิทธิภาพสูงสุด ให้ค่าความถูกต้อง เท่ากับ 97.85% ความระลึก 99.38% รองลงมาคือวิธีการเพื่อนบ้านใกล้ที่สุดที่ให้ค่าความถูกต้องเท่ากับ 97.36% และวิธีแบบเบย์ ให้ค่าความถูกต้องเท่ากับ 96.75% จากผลการเปรียบเทียบประสิทธิภาพครั้งนี้ สามารถนำวิธีต้นไม้ตัดสินใจ ใช้เป็นตัวแบบในการทำนายความเสี่ยงในการติดเชื้อโควิด-19

คำสำคัญ: เหมืองข้อมูล วิธีต้นไม้ตัดสินใจ วิธีแบบเบย์ โควิด-19 วิธีการเพื่อนบ้านใกล้ที่สุด

A Comparison of the COVID-19 Risk Prediction Model with Data Mining Techniques

Jirayu Sitichai[1*] and Phanaya Palvisut[1]

[1]Data Science, Faculty of Science and Technology, Nakhon Pathom Rajabhat University
*Jirayu7991@gmail.com

Abstract

The purpose of this study was to compare the effectiveness of models in predicting the likelihood of contracting COVID-19. Three data mining techniques were used, namely, decision tree method Naive Bayes method and K-Nearest Neighbor method The efficacy of appropriate identification classification models for predicting the likelihood of contracting COVID-19 was compared using 5,434 rows of 21 columns of COVID-19 data from the kaggle website. Data were analyzed on the basis of CRISP-DM method. Use RapidMiner. using Rapidminer. in modeling in the analysis of data, accuracy, precision, recall.

The results showed that the decision tree method was the most effective. The accuracy was 97.85%, followed by the K-Nearest Neighbor method, which gave an accuracy of 97.36%, and Naive Bayes method gave an accuracy of 97.36%. 96.75% from the results of this performance comparison. can adopt the decision tree method can be used as a model for predicting the risk of contracting COVID-19.

Keywords: Data mining, Decision tree method, Naive Bayes method, COVID-19, K-Nearest Neighbor method

1.บทนำ

ไวรัสโคโรนา (Coronavirus) [1] หรือ COVID-19 เป็นไวรัสที่ถูกพบครั้งแรกในปี 1960 แต่ยังไม่ทราบแหล่งที่มาอย่างชัดเจนว่ามาจากที่ใด แต่เป็นไวรัสที่สามารถติดเชื้อได้ทั้งในมนุษย์และสัตว์ ซึ่งสามารถกลายพันธุ์ได้โดยการแพร่เชื้อจากคนสู่คนและสัตว์สู่สัตว์ในลักษณะเดียวกับไข้หวัดใหญ่โดยผ่านการติดเชื้อจากการไอหรือจาม โดยหายใจเอาฝอยละอองจากผู้ป่วยเข้าไป หรือจากการเอามือไปจับพื้นผิวที่มีฝอยละอองเหล่านั้นแล้วมาจับตามใบหน้า ปัจจุบันทางการแพทย์ยังไม่สามารถผลิตวัคซีนหรือยาต้านไวรัสได้ร้อยละ 100 เนื่องจากไวรัสมีการกลายพันธุ์ ทำได้แค่รักษาผู้ที่ติดเชื้อและเฝ้าระวังผู้ที่มีความเสี่ยงในการติดเชื้อ ประเทศอื่นๆได้เห็นถึงความรุนแรงของการกลายพันธุ์ ทำให้ไวรัสแพร่ระบาดอย่างรวดเร็ว ในครั้งนี้จึงได้ออกมาตรการในการ ควบคุมดูแลและปิดประเทศป้องกันการแพร่ระบาดของไวรัสโคโรนา

ปัจจุบันประเทศไทยมีผู้ติดเชื้อและผู้ที่เสียชีวิตค่อนข้างสูง ในปี 2022 และมีแนวโน้มที่จะมีผู้ติดเชื้อเพิ่มขึ้นเนื่องจากการกลายพันธุ์ของไวรัสก่อให้เกิดสายพันธุ์ใหม่ ซึ่งสายพันธุ์ที่กำลังระบาดในปัจจุบันคือ โอไมครอน(Omicron) โดยมีการกลายพันธุ์ของยีนมากถึง 50 กว่าตำแหน่ง โดย 32 ตำแหน่งเกิดขึ้นบนโปรตีนหนามแหลม หรือที่เรียกว่า Spike Protein ซึ่งเป็นโปรตีนที่ไวรัสใช้

ในการเข้าสู่เซลล์ของร่างกายมนุษย์ ซึ่งพบมากกว่าทุกสายพันธุ์ และมากกว่าสายพันธุ์เดลตา ถึง 2 เท่า และพบการกลายพันธุ์ที่ส่วนตัวรับ ซึ่งไวรัสใช้จับยึดกับเซลล์ของคนถึง 10 ตำแหน่ง [2]

การนำเทคโนโลยีเหมืองข้อมูล (Data Mining) [3] มาช่วยในการทำนายเพื่อการวางแผนให้ตรงกับความต้องการใช้งาน เป็นการนำความรู้จากข้อมูลที่มีอยู่มาใช้ในการทำนายข้อมูลใหม่ที่จะเกิดขึ้นในอนาคต ซึ่งการทำเหมืองข้อมูลคือการกระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น ในปัจจุบันการทำเหมืองข้อมูลได้ถูกนำมาประยุกต์ใช้งานหลายประเภท ทั้งในด้านธุรกิจที่ช่วยในการตัดสินใจของผู้บริหาร ในด้านวิทยาศาสตร์และแพทย์ รวมทั้งในด้านเศรษฐกิจและสังคม

ในการวิจัยครั้งนี้ ได้นำข้อมูลของโควิด-19 มาวิเคราะห์เพื่อพัฒนาหาโมเดลที่สามารถทำนายหาค่าความแม่นยำที่จะหาผู้ที่เสี่ยงติดเชื้อโควิด-19

2. วัตถุประสงค์การวิจัย

1) เพื่อศึกษาตัวแบบในการทำนายความเสี่ยงในการติดเชื้อโควิด-19

2) เพื่อเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ เทคนิคแบบเบย์และวิธีการเพื่อนบ้านใกล้ที่สุด

3. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ทฤษฎีที่เกี่ยวข้อง

เหมืองข้อมูล

เหมืองข้อมูล [4] คือการค้นหาหรือการสกัดความรู้จากฐานข้อมูลขนาดใหญ่ กระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบแนวทางและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น โดยอาศัยหลักการทางคณิตศาสตร์ สถิติ เพื่อนำความรู้ที่ได้นั้นมาใช้ในการแก้ปัญหา วางแผน หรือการดำเนินกลยุทธ์ขององค์กรให้ประสบความสำเร็จสูงสุด ซึ่งการวิเคราะห์ข้อมูลด้วยเทคนิคเหมืองข้อมูลสามารถแบ่งได้เป็น 2 ประเภทหลักๆ คือ

1.เทคนิคการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) [5] จะเน้นที่การพิจารณาข้อมูลเป็นหลัก เช่น พิจารณาว่าข้อมูลมีความสัมพันธ์กันในลักษณะใดบ้าง เทคนิคในประเภทนี้จะแบ่งย่อยได้อีกคือ เทคนิคการค้นหากฎความสัมพันธ์ (Association Rule) และการแบ่งกลุ่มข้อมูล (Clustering)

2.เทคนิคการเรียนรู้แบบมีผู้สอน (Supervised Learning) [6] เน้นการเรียนรู้จากข้อมูลที่มีอยู่ในอดีตเพื่อนำมาสร้างโมเดลสำหรับทำนายหรือคาดการณ์สิ่งที่จะเกิดขึ้นในอนาคต สามารถแบ่งย่อยได้อีกคือ การจำแนกประเภทข้อมูล (Classification) และการประมาณค่าข้อมูล (Regression) ซึ่งทั้งสองเทคนิคจะมีความแตกต่างกันที่คำตอบที่ต้องการทำนายซึ่งการจำแนกประเภทข้อมูล จะทำนายข้อมูลที่มีค่าน้อย เช่น เพศชาย เพศหญิง หรือค่าที่ไม่ใช่ตัวเลขนั้นเอง ส่วนการประมาณค่าข้อมูล จะใช้กับข้อมูลคำตอบที่เป็นตัวเลขเท่านั้น

ในการวิจัยครั้งนี้ ได้ใช้เทคนิคการเรียนรู้แบบมีผู้สอน (Supervised Learning) แบบการจำแนกประเภทข้อมูล โดยใช้อัลกอริทึมเพื่อใช้ทำนายแนวโน้มให้แก่ข้อมูลที่ใช้ทำนาย

วิธีต้นไม้ตัดสินใจ [7]

เป็นการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ในรูปแบบโครงสร้างต้นไม้ สามารถสร้างแบบจำลองการจัดหมวดหมู่ได้จากกลุ่มตัวอย่างข้อมูลที่กำหนดไว้ล่วงหน้า และพยากรณ์กลุ่มของรายการที่ยังไม่เคยนำมาจัดหมวดหมู่ได้ด้วยรูปแบบของ ต้นไม้ (Tree) คลังข้อมูลและความรู้ระบบสุขภาพ

วิธีแบบเบย์ [8]

เป็นการทำเหมืองข้อมูลที่ถูกสร้างขึ้นโดยหลักความน่าจะเป็น ซึ่งจะใช้การวิเคราะห์หาความน่าจะเป็นของสิ่งที่ยังไม่เคยเกิดขึ้นด้วยการคาดเดาจากสิ่งที่เคยเกิดขึ้นมาก่อน โดยใช้ทฤษฎีของ Thomas Bayes ในการแก้ปัญหา

วิธีการเพื่อนบ้านใกล้ที่สุด [9]

เป็นวิธีการจำแนกประเภทข้อมูลวิธีหนึ่ง โดยจัดว่าเป็นการจำแนกแบบมีผู้ฝึกสอน (Supervised Machine Learning Algorithm) หรือ ทราบคำตอบอยู่แล้ว จากนั้นจะใช้โมเดลในการจำแนกประเภท ข้อมูลจากข้อมูลที่รู้คำตอบ LSONGKIAT

4. งานวิจัยที่เกี่ยวข้อง

สุรวัชร และสายชล สิ [10] ได้ทำการเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มการเป็นโรคไตเรื้อรัง ซึ่งเป็นงานวิจัยเกี่ยวการเป็นโรคไตเรื้อรัง งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่ม โดยเลือกใช้วิธีความใกล้เคียงกันมากที่สุด วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีฐานกฎ และวิธีแบบเบย์ ผลที่ได้คือวิธีการจำแนกกลุ่มที่มีประสิทธิภาพการจำแนกดีที่สุดคือ วิธีต้นไม้ตัดสินใจ ซึ่งให้ค่าความถูกต้อง คือ 100% และค่าความคลาดเคลื่อนกำลังสองเฉลี่ยคือ 0.0059

รักถิ่น [11] ได้ทำการพยากรณ์ความเสี่ยงการเกิดโรคเบาหวานโดยเทคนิคการทำเหมืองข้อมูล: กรณีโรงพยาบาลมหาสารคาม โดยงานวิจัยนี้ มีวัตถุประสงค์เพื่อ (1) เพื่อศึกษาการพยากรณ์ความเสี่ยงการเกิดโรคเบาหวาน โดยนำค่าปัจจัยเสี่ยงของผู้ป่วยโรคเบาหวาน มาใช้ในการพยากรณ์ (2) เพื่อหาค่าประสิทธิภาพของตัวแบบการพยากรณ์ มีวิธีดำเนินการวิจัยตามกระบวนการของ ของคริสป์-ดีเอ็ม ประกอบด้วย 6 ขั้นตอน คือขั้นตอนความเข้าใจในธุรกิจ ขั้นตอนความเข้าใจข้อมูล ขั้นตอนการเตรียมข้อมูล ขั้นตอนการจัดทำแบบจำลองเหมืองข้อมูล ขั้นตอนการประเมินผล และขั้นตอนการนำแบบจำลองไปใช้งาน ผลการวิจัยพบว่า (1) ได้ตัวแบบการพยากรณ์ความเสี่ยงการเกิดโรคเบาหวาน ด้วยอักอริทึมต้นไม่ตัดสินใจแบบ ID3 (2) ผลการประเมินประสิทธิภาพตัวแบบ จากการแบ่งข้อมูลทดสอบออกเป็น 5 ชุด ที่ค่าความถูกต้อง ได้ 69.45%

เอพร และคณะ [12] การศึกษาสภาวะเสี่ยงโรคของผู้สูงอายุด้วยเทคนิคเหมืองข้อมูล เป็นงานวิจัยเกี่ยวกับปัญหาในศึกษาออกกลางคันในเวลาเรียน โดยงานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาความสัมพันธ์ของปัจจัยที่ส่งผลต่อการเกิดโรค และรูปแบบสภาวะเสี่ยงโรคของผู้สูงอายุด้วยเทคนิคเหมืองข้อมูล โดยใช้ข้อมูลการคัดกรองสุขภาพของผู้สูงอายุ อำเภอเมืองสุราษฎร์ธานี จังหวัดสุราษฎร์ธานี ปี พ.ศ.2561 จำนวน 3,875 เรคอร์ด โดยแบ่งการศึกษาออกเป็น 2 ประเด็น คือ ศึกษาความสัมพันธ์ของปัจจัยที่ส่งผลต่อการเกิดโรค ด้วยเทคนิค Association Rule และศึกษารูปแบบสภาวะเสี่ยงโรคของผู้สูงอายุ ด้วยการแบ่งกลุ่ม 3 กลุ่ม คือ กลุ่มปกติ กลุ่มเสี่ยง และกลุ่มป่วย ด้วยเทคนิค Classification ในการคัดกรองสุขภาพเบื้องต้นของ ผลการวิจัยพบว่าการศึกษาความสัมพันธ์ของการเป็นโรคต่าง ๆ โดยใช้เทคนิคอัลกอริทึมในกลุ่ม Decision Tree 3 โมเดลได้แก่ โมเดล C4.5 โมเดล Partial Rule และโมเดล Induction ผลการวิจัยพบว่า เทคนิค Decision Tree J48 ให้ค่าความถูกต้องมากที่สุด โดยมีค่าความ

แม่นยำ (Correctly) ร้อยละ 99.796 ค่าความถูกต้อง (Precision) ร้อยละ 0.998 ค่าระลึก (Recall) ร้อยละ 0.998 และค่าความเหวี่ยง (F-measure) ร้อยละ 0.998

เพชรรัตน์ และคณะ [13] ได้ทำตัวแบบประเมินภาวะความเสี่ยงการเป็นโรคซึมเศร้าของนักศึกษาด้วยเทคนิคเหมืองข้อมูล การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อ 1) ศึกษาและพัฒนาตัวแบบประเมินภาวะความเสี่ยงการเป็นโรคซึมเศร้าของนักศึกษาด้วยเทคนิคการเหมืองข้อมูล กลุ่มตัวอย่างที่ใช้ในการวิจัย ได้แก่ นักศึกษาระดับปริญญาตรี มหาวิทยาลัยราชภัฏนครปฐม จำนวน 600 คน โดยสุ่มตัวอย่างจากทุกคณะ เครื่องมือที่ใช้ในการวิจัย ได้แก่ 1) แบบสอบถามข้อมูลทั่วไป และ 2) แบบประเมินแบบทดสอบภาวะซึมเศร้า PHQ-9 เทคนิคการวิเคราะห์ข้อมูลที่ใช้ในการวิจัย คือ เทคนิคเหมืองข้อมูล จำนวน 5 อัลกอริทึม ได้แก่ Random, Tree, LMT, PART J48, and JRIP และทำการเปรียบเทียบประสิทธิภาพด้วยการทดสอบ Cross-validation ผลการวิจัยพบว่า Random Tree มีประสิทธิภาพสูงสุดโดยมีค่าความถูกต้อง 96.00% รองลงมาคือ อัลกอริทึม LMT มีค่า 95.90% อัลกอริทึม PART มีค่า 95.10% อัลกอริทึม J48 มีค่า 94.80% และอัลกอริทึม JRIP มีค่า 94.00%

วนิดา และคณะ [14] ได้ทำการพัฒนาแบบจำลองปัจจัยที่มีผลต่อการเป็นโรคเบาหวานด้วยเทคนิคต้นไม้ตัดสินใจ มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองของปัจจัยที่มีผลต่อการเป็นโรคเบาหวานด้วยเทคนิคเหมืองข้อมูลแบบต้นไม้ตัดสินใจเพื่อช่วยในการวิเคราะห์หาแบบจำลองของปัจจัยที่มีผลต่อการเป็นโรคเบาหวาน ปัจจัยเสี่ยงที่มีผลต่อการเกิดโรคเบาหวานถือเป็นสารสนเทศที่มีความสำคัญที่จะช่วยให้หน่วยงานทางด้านการแพทย์นำไปใช้สำหรับวางแผนกลยุทธ์ในการป้องกันโรคที่ตรงกับกลุ่มเป้าหมาย ในการดำเนินงานวิจัยใช้ข้อมูลผู้เข้ารับบริการที่โรงพยาบาลด่านขุนทด จังหวัดนครราชสีมา ระหว่างปี 2550 – 2555 จำนวนทั้งสิ้น 4,402 ราย แบ่งข้อมูลสำหรับฝึกและทดสอบแบบจำลองออกเป็นร้อยละ 90:10 ตามลำดับ พัฒนาแบบจำลองด้วยอัลกอริทึม J48 ซึ่งเป็นเทคนิคแบบต้นไม้ตัดสินใจ ประเมินประสิทธิภาพของแบบจำลองด้วยค่าความแม่นตรง ผลการวิจัยพบว่าแบบจำลองที่พัฒนาให้ประสิทธิภาพที่มีค่าความแม่นตรง 76.14% และสามารถสร้างกฎการจำแนกจากต้นไม้ตัดสินใจทั้งสิ้น 97 กฎ ซึ่งพบว่าปัจจัยเสี่ยงที่อาจก่อให้เกิดโรคเบาหวาน ได้แก่ อายุ เพศ สถานะภาพ ที่อยู่ อาชีพ ประวัติความดันโลหิตเกินมาตรฐาน ประวัติค่าดัชนีมวลกายเกินมาตรฐาน พฤติกรรมการสูบบุหรี่ พฤติกรรมการดื่มสุรา และประวัติครอบครัวเป็นเบาหวาน

5. วิธีดำเนินการวิจัย

CRISP-DM เป็นกระบวนการในการวิเคราะห์ที่ใช้กันอย่างแพร่หลายซึ่งประกอบด้วย 6 ขั้นตอน ได้แก่

1. Business Understanding เป็นขั้นตอนการทำความเข้าใจโจทย์โดยการวิเคราะห์ข้อมูล โดยข้อมูลที่นำมาวิเคราะห์มาทำการทำนายความเสี่ยง เพื่อเป็นตัวช่วยในการคาดการณ์ความเสี่ยงในการติดเชื้อโควิด-19

2. Data Understanding เป็นขั้นตอนการดูว่ามีข้อมูลอะไรบ้างที่สามารถนำมาใช้ในการวิเคราะห์ของโปรเจคนี้ได้บ้าง จากการศึกษาและวิเคราะห์ปัญหาที่เกี่ยวข้องกับงานวิจัย และการวิเคราะห์ข้อมูล โดยวิเคราะห์จากข้อมูลที่เลือกมา เป็นไฟล์ Excel จำนวนข้อมูลทั้งหมดมี 5,434 แถว 21 คอลัมน์ ดังนี้

ตารางที่ 1 ข้อมูลที่ใช้วิเคราะห์

คุณลักษณะ	ข้อมูล	ความหมายข้อมูล
1) Breathing Problem ปัญหาการหายใจ	- Yes - No	- มีอาการ - ไม่มีอาการ
2) Fever ไข้	- Yes - No	- มีอาการ - ไม่มีอาการ

คุณลักษณะ	ข้อมูล	ความหมายข้อมูล
3) Dry Cough อาการไอแห้ง	- Yes - No	- มีอาการ - ไม่มีอาการ
4) Sore Throat เจ็บคอ	- Yes - No	- มีอาการ - ไม่มีอาการ
5) Running Nose วิ่งจมูก	- Yes - No	- มีอาการ - ไม่มีอาการ
6) Asthma หอบหืด	- Yes - No	- มีอาการ - ไม่มีอาการ
7) Chronic Lung Disease โรคปอดเรื้อรัง	- Yes - No	- เป็น - ไม่เป็น
8) Headache ปวดหัว	- Yes - No	- มีอาการ - ไม่มีอาการ
9) Heart Disease โรคหัวใจ	- Yes - No	- เป็น - ไม่เป็น
10) Diabetes เบาหวาน	- Yes - No	- มีอาการ - ไม่มีอาการ
11) Hyper Tension ความตึงเครียดสูง	- Yes - No	- มีอาการ - ไม่มีอาการ
12) Fatigue ความเหนื่อยล้า	- Yes - No	- มีอาการ - ไม่มีอาการ
13) Gastrointestinal ระบบทางเดินอาหาร	- Yes - No	- มีอาการ - ไม่มีอาการ
14) Abroad Travel การเดินทางไปต่างประเทศ	- Yes - No	- เดินทาง - ไม่เดินทาง
15) Contact with COVID Patient การติดต่อกับผู้ป่วยโควิด	- Yes - No	- ติดต่อ - ไม่ติดต่อ
16) Attended Large Gathering เข้าร่วมชุมนุมใหญ่	- Yes - No	- ใช่ - ไม่ใช่
17) Visited Public Exposed Places เยี่ยมชมสถานที่สาธารณะที่เปิดเผย	- Yes - No	- ใช่ - ไม่ใช่

คุณลักษณะ	ข้อมูล	ความหมายข้อมูล
18) Family working in Public Exposed Places ครอบครัวที่ทำงานในสถานที่สาธารณะ	- Yes - No	- ใช่ - ไม่ใช่
19) Wearing Masks สวมหน้ากาก	- Yes - No	- สวมใส่ - ไม่สวมใช่
20) Sanitization from Market	- Yes - No	- ใช่ - ไม่ใช่
21) COVID-19 โควิด-19	- Yes - No	- มีอาการ - ไม่มีอาการ

3. Data Preparation เป็นขั้นตอนของการเตรียมข้อมูลโดยอาจจะเป็นการเชื่อมโยงข้อมูล (Join) สำหรับการวิเคราะห์ข้อมูลในงานวิจัยครั้งนี้ด้วยโปรแกรม Rapidminer

การแปลงข้อมูล หมายถึง การเปลี่ยนสภาพของข้อมูลที่ศึกษาให้มีการแจกแจงแบบปกติหรือทำให้ความแปรปรวนมีค่าเท่ากัน เนื่องจากข้อตกลง เบื้องต้นของการทดสอบสถิติบางตัวได้กำหนดไว้ เช่น การทดสอบค่าเฉลี่ย การทดสอบความแปรปรวน (Analysis of Variance) การวิเคราะห์การถดถอย (Regression Analysis)

4. Modeling เป็นขั้นตอนการวิเคราะห์ข้อมูลโดยการใช้เทคนิคต่างๆ ของเหมืองข้อมูล

การจำแนกประเภทข้อมูล เป็นขั้นตอนการทดสอบการจำแนกประเภทข้อมูล หลังจากทำการเลือกคุณลักษณะของข้อมูลที่มีความเหมาะสมและสอดคล้องมากที่สุด โดยทำการเปรียบเทียบค่าเฉลี่ยความถูกต้องการจำแนกประเภทข้อมูลของแต่ละกลุ่มการทดสอบ โดยใช้เครื่องมือ Rapidminer และ Excel ซึ่งกลุ่มตัวอย่าง คือ https://www.kaggle.com/datasets/hemanthhari/symptoms-and-covid-presence โดยใช้เทคนิค 3 แบบ วิธีต้นไม้ตัดสินใจ วิธีแบบเบย์ วิธีการเพื่อนบ้านใกล้ที่สุด

5. Evaluation การประเมินผล เป็นกระบวนการในการตัดสินคุณค่าให้กับสิ่งต่าง ๆ โดยนำผลที่ได้จากการวัดมาเทียบกับเกณฑ์ที่กำหนดไว้ แล้วทำการพิจารณาตัดสินว่าสิ่งนั้นมีคุณภาพในระดับใด เช่น ดี พอใช้ ไม่ดี

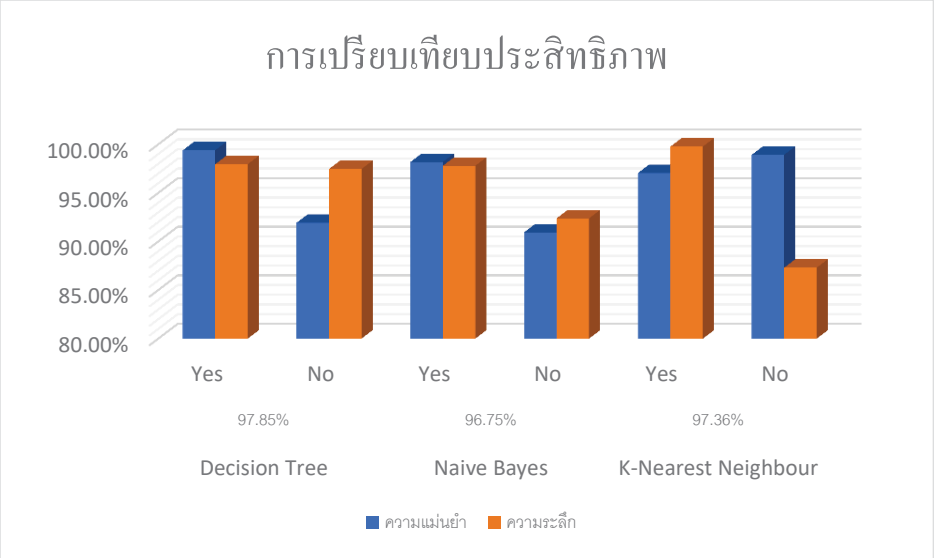
6. Deployment เป็นขั้นตอนการนำผลลัพธ์การวิเคราะห์ที่ได้ไปใช้งานต่อไป เช่น การนำเสนอข้อมูลที่เป็นประโยชน์ในด้านการแพทย์ เพื่อใช้ในการตัดสินใจของหน่วยงานที่ให้บริการทางการแพทย์

6. ผลการวิจัย

ผู้วิจัยได้เปรียบเทียบประสิทธิภาพเทคนิคการเลือกคุณลักษณะที่เหมาะสมสำหรับการจำแนกประเภทข้อมูลผลการเปรียบเทียบการคัดเลือกคุณลักษณะในการคัดเลือกคุณลักษณะด้วยเทคนิคแบบ คือ วิธีต้นไม้ตัดสินใจ วิธีแบบเบย์ วิธีการเพื่อนบ้านใกล้ที่สุด ดังนี้

ตารางที่ 2 การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลทั้ง 3 เทคนิค

Algorithm	Accuracy	Class	Precision	Recall
Decision Tree	97.85%	Yes	99.38%	97.95%
		No	91.92%	97.46%
Naive Bayes	96.75%	Yes	98.17%	97.79%
		No	90.94%	92.38%
K-Nearest Neighbour	97.36%	Yes	97.04%	99.77%
		No	98.92%	87.30%



ภาพที่ 1 การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูล

จากตารางที่ 2 ผลการเปรียบเทียบประสิทธิภาพ พบว่า วิธีต้นไม้ตัดสินใจ ให้ประสิทธิภาพที่ดีที่สุด โดยให้ค่าความถูกต้อง 97.85% ค่าความแม่นยำที่จะเป็น 99.38% ไม่เป็น 91.92% ค่าความระลึกที่จะเป็น 97.95% ไม่เป็น 97.46% และเมื่อเปรียบเทียบค่า 3 ค่า ทั้ง 3 เทคนิค ทำให้ทราบว่าทุกๆค่ามีความใกล้เคียงกัน อันดับที่ 1 วิธีต้นไม้ตัดสินใจ อันดับที่ 2 วิธีการเพื่อนบ้านใกล้ที่สุด และอันดับที่ 3 วิธีแบบเบย์

7. สรุป อภิปรายผลการวิจัย

ผลการวิจัย ครั้งนี้โดยพัฒนาและเปรียบเทียบตัวแบบการจำแนกทั้ง 3 เทคนิค ได้แก่ วิธีต้นไม้ตัดสินใจ วิธีแบบเบย์ วิธีการเพื่อนบ้านใกล้ที่สุด ซึ่งผลการประเมินประสิทธิภาพตัวแบบ คือ วิธีต้นไม้ตัดสินใจ ซึ่งได้ค่าที่สูงที่สุดจากข้อมูล ค่าความถูก

ต้องได้ 97.85% ค่าความระลึกที่จะเป็น 97.95% ไม่เป็น 97.46% จึงสรุปได้ว่า วิธีต้นไม้ตัดสินใจ เป็นตัวแบบที่เหมาะสมที่สุดที่จะนำมาใช้จำแนกข้อมูล ผลการทดลองใช้โปรแกรม RapidMiner พบว่า วิธีต้นไม้ตัดสินใจ มีความถูกต้องมากที่สุด

8. ข้อเสนอแนะ

การวิจัยครั้งนี้เป็นการวิจัยที่ใช้โปรแกรม RapidMiner เพื่อนำมาประยุกต์ใช้ในการจำแนกข้อมูล โดยที่มีเครื่องมือต่างๆ ให้เลือกใช้ อาทิเช่น วิธีต้นไม้ตัดสินใจ วิธีแบบเบย์ วิธีการเพื่อนบ้านใกล้ที่สุด โดยสามารถนำเครื่องมือเหล่านี้มาเปรียบเทียบการจำแนกเพื่อหาค่าความถูกต้องมากที่สุด

9. เอกสารอ้างอิง

[1] มหาวิทยาลัยมหิดล คณะแพทยศาสตร์ศิริราชพยาบาล ศูนย์การแพทย์กาญจนาภิเษก. (2563). **โควิด-19 คืออะไร**. ค้นเมื่อ [20 เมษายน 2565] จาก https://www.gj.mahidol.ac.th/main/knowledge-2/covid19is/.

[2] กรุงเทพประกันชีวิต. (2565). **เรื่องควรรู้เกี่ยวกับโรคโควิด 19 (COVID19 : SARS-CoV-2)**. ค้นเมื่อ [20 เมษายน 2565] จาก https://www.bangkoklife.com/en/articles/49/87.

[3] ศจี วานิช. (2558). Data Mining (เหมืองข้อมูล). ค้นเมื่อ [20 เมษายน 2565] จาก http://sajeegm301.blogspot.com/2015/11/data-mining.html.

[4] กองโลจิสติกส์ กรมส่งเสริมอุตสาหกรรรม. (2562). **การทำเหมืองข้อมูล (Data Mining)**. ค้นเมื่อ [20 เมษายน 2565] จาก https://dol.dip.go.th/th/category/2019-02-08-08-57-30/2019-03-15-08-49-57.

[5] มูลนิธิโครงการสารานุกรมไทยสำหรับเยาวชน. (2544). **การเรียนรู้แบบไม่มีผู้สอน (Un-supervised Learning)**.ค้นเมื่อ [20 เมษายน 2565] จาก https://www.saranukromthai.or.th/sub/book/book.php?book=25&chap=5&page=t25-5-infodetail06.html.

[6] GlurGeek.Com. (2562). **Supervised Learning (การเรียนรู้แบบมีผู้สอน) คืออะไร**. ค้นเมื่อ [20 เมษายน 2565] จาก https://www.glurgeek.com/education/ie311supervisedlearning/.

[7] คลังข้อมูลและความรู้ระบบสุขภาพ สถาบันวิจัยระบบสาธารณสุข (สวรส.). (2563). **ต้นไม้การตัดสินใจ**. ค้นเมื่อ [20 เมษายน 2565] จาก https://kb.hsri.or.th/dspace/handle/11228/2964?locale-attribute=th

[8] Knowledge. (2562). **Naive Bayes classification #1**. ค้นเมื่อ [20 เมษายน 2565] จาก http://cakeknowledgeblogs.blogspot.com/2019/08/naive-bayes-classification-1.html.

[9] SKLSONGKIAT OUTSOURCE SERVICES. (2564). **ทำไมต้องใช้ K-Nearest Neighbor (K-NN)**. ค้นเมื่อ [20 เมษายน 2565] จาก https://www.sklsongkiat.com/articles/detail/ทำไมต้องใช้-k-nearest-neighbor-k-nn.

[10] สุรวัชร ศรีเปารยะ และสายชล สินสมบูรณ์ทอง. (2560). **การเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มการเป็นโรคไตเรื้อรัง** ภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.

[11] รักถิ่น เหลาหา. (2560). **การเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มการเป็นโรคไตเรื้อรัง ซึ่งเป็นงานวิจัยเกี่ยวการเป็นโรคไตเรื้อรัง**. Mojor Management Information System Faculty of Science and Technology Rajabhat Maha Sarakham University.

[12] เอพร โมฬี นิธิศ เสาแก้วและบุษยมาศ เหมณี. (2562). **การศึกษาสภาวะเสี่ยงโรคของผู้สูงอายุด้วยเทคนิคเหมืองข้อมูล เป็นงานวิจัยเกี่ยวกับปัญหาในศึกษาออกกลางคันในเวลาเรียน** คณะมนุษยศาสตร์และสังคมศาสตร์ มหาวิทยาลัยราชภัฏสุราษฎร์ธานี คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏสุราษฎร์ธานี.

[13] เพชรรัตน์ ม่วงน้อย จักรพันธ์ พลาผลและภรัณยา ปาลวิสุทธิ์. (2564). **ตัวแบบประเมินภาวะความเสี่ยงการเป็นโรคซึมเศร้าของนักศึกษาด้วยเทคนิคเหมืองข้อมูล** สาขาวิชาเทคโนโลยีสารสนเทศ สาขาวิชาวิทยาการข้อมูล คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม**.**

[14] วนิดา พงษ์สงวน ทิพยา ถินสูงเนินและมาโนช ถินสูงเนิน. (2562). **การพัฒนาแบบจำลองปัจจัยที่มีผลต่อการเป็นโรคเบาหวานด้วยเทคนิคต้นไม้ตัดสินใจ** สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครราชสีมา**.**