การเปรียบเทียบประสิทธิภาพของแบบจำลองการทำนายความเสี่ยงโรคมะเร็งปอดด้วยเทคนิคเหมืองข้อมูล

Efficiency Comparison of Lung Cancer Risk Prediction Models using Data-mining Techniques

ธวัชชัย เหล็กดี*

Thawatchai Lekdee*

กองคุ้มครองและส่งเสริมภูมิปัญญาการแพทย์แผนไทยและแพทย์พื้นบ้านไทย กรมการแพทย์แผนไทยและการแพทย์ทางเลือก

Division of Protection and Promotion of Thai Traditional and Indigenous Medicine,

Department of Thai Traditional Medicine and Alternative Medicine

รัฐพรรณ สันติอโนทัย

Ruthaphan Santianotai

ภาควิชาสาธารณสุขศาสตร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏเชียงใหม่

Department of Public Health, Faculty of Science and Technology, Chaing Mai Rajabhat University

เจษฎา อุดมพิทยาสรรพ์

Jadsada Udompittayason

วิทยาลัยการสาธารณสุขสิรินธร

Sirindhorn College of Public Health

E-mail : doctor.aoteza@gmail.com*, ruthaphan_san@g.cmru.ac.th and jadsada@scphtrang.ac.th

*Corresponding author

(Received: 4 November 2023, Revised: 22 December 2023, Accepted: 25 December 2023)

https://doi.org/10.57260/stc.2024.705

บทคัดย่อ

การวิจัยนี้มีวัตถุประสงค์เพื่อสร้างและเปรียบเทียบประสิทธิภาพของแบบจำลองที่ใช้สำหรับทำนายความเสี่ยงโรคมะเร็งปอด โดยวิเคราะห์ตามกระบวนการมาตรฐานของการทำเหมืองข้อมูล (CRISP-DM) ข้อมูลทั้งหมดมี 310 รายการ แบ่งเป็นสองกลุ่ม กลุ่มที่เป็นโรคมะเร็งปอด 270 รายการ และกลุ่มที่ไม่เป็นโรคมะเร็งปอด 39 รายการ ซึ่งถูกเรียกว่า คลาส YES และ คลาส No ตามลำดับ สมดุลข้อมูลด้วยวิธีการสังเคราะห์ข้อมูลเพิ่ม (Synthetic minority oversampling technique: SMOTE) และสร้างแบบจำลองใช้เทคนิคการทำเหมืองข้อมูล 4 เทคนิค ประกอบด้วย เทคนิคต้นไม้ตัดสินใจ เทคนิคป่าสุ่ม เทคนิคนาอีฟเบย์ และเทคนิคซัพพอรตเวกเตอรแมชชีน และเปรียบเทียบประสิทธิภาพแบบจำลองด้วยค่าความถูกต้อง

(Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าประสิทธิภาพโดยรวม (F-measure) โดยใช้โปรแกรม RapidMiner studio version 10.1 ในการสร้างแบบจำลองและการวิเคราะห์ข้อมูล ผลการวิจัยพบว่า เทคนิคป่าสุ่มเป็นเทคนิคที่ดีที่สุด ให้ค่าความถูกต้อง 94.63% ค่าความแม่นยำ 92.92% ค่าความระลึก 96.67% และค่าประสิทธิภาพโดยรวม 94.73% ซึ่งผลการวิจัยนี้สามารถนำไปสร้างเป็นระบบสารสนเทศเพื่อพยากรณ์ผู้ป่วยมะเร็งปอด โดยเป็นการคัดกรองข้อมูลผู้ป่วยเบื้องต้นก่อนถึงมือแพทย์

คำสำคัญ: ต้นไม้ตัดสินใจ ป่าสุ่ม นาอีฟเบย์ โรคมะเร็งปอด

Abstract

This research aims to create and compare the efficiency of models used for predicting the risk of lung cancer by analyzing according to the Cross-Industry Standard Process for Data Mining (CRISP-DM). The dataset comprises 310 items, divided into two groups: 270 instances with lung cancer (Class YES) and 39 instances without lung cancer (Class NO). Data balance was achieved using the Synthetic Minority Oversampling Technique (SMOTE). Four data mining techniques were employed: Decision Tree, Random Forest, and Naïve Bayes, as well as Support Vector Machine. Model performance was evaluated using metrics such as Accuracy, precision, recall, and F-measure. RapidMiner Studio Version 10.1 was utilized for model creation and data analysis. The findings reveal that the Random Forest technique outperformed others, yielding an accuracy of 94.63%, precision of 92.92%, recall of 96.67%, and an overall F-measure of 94.73%. This research suggests that the Random Forest technique is the most effective for predicting lung cancer risk, providing valuable insights for potential integration into an information system for preliminary patient screening before reaching medical professionals.

Keywords: Decision tree, Random forest, Naïve bayes, Lung cancer disease

บทนำ

ในประเทศไทยมีผู้เสียชีวิตมาจากโรคต่าง ๆ ซึ่งโรคมะเร็งเป็นสาเหตุการเสียชีวิตอันดับหนึ่ง โดยที่โรคมะเร็งถือเป็นโรคที่มีความผิดปกติของเซลล์ในอวัยวะต่าง ๆ ของร่างกาย โดยเกิดการเปลี่ยนแปลงทางพันธุกรรมของเซลล์ก่อให้เกิดเป็นเซลล์มะเร็งที่มีการเจริญเติบโตโดยไม่อยู่ภายใต้การควบคุมที่เหมาะสม ทำให้เกิดเป็นก้อนเนื้อมะเร็งที่เติบโตรบกวนการทำงานของเซลล์ปกติในอวัยวะ นอกจากนี้ยังสามารถลุกลามแพร่กระจายไปยังอวัยวะอื่นได้ (จีราพร บวรอารักษ์ และคณะ, 2562)

โรคมะเร็งปอดเป็นมะเร็งที่พบได้บ่อยและเป็นสาเหตุการเสียชีวิตอันดับต้นของประชากรโลกจากสถิติมะเร็งของไทย ในช่วงปี พ.ศ. 2559 ถึง 2561 พบว่าในเพศชายมะเร็งปอดพบสูงเป็นอันดับสอง (Rojanamatin et al., 2021) และปัจจัยเสี่ยงต่อการเกิดโรคมะเร็งปอดมาจากหลายสาเหตุ เช่น การสูบบุหรี่ สิ่งแวดล้อมที่อยู่อาศัย สิ่งแวดล้อมที่ทำงาน และควันจากการเผาไหม้น้ำมัน และถ่านหิน เป็นต้น (รักถิ่น เหลาหา, 2553) นอกจากนี้ผู้ที่สูบบุหรี่ ที่เป็นโรคปอดอุดกั้นเรื้อรังจะยิ่งเพิ่มปัจจัยเสี่ยงต่อมะเร็งปอด สารก่อมะเร็งที่อาจเป็นสาเหตุของโรคในผู้ป่วย 10-15% ซึ่งไม่สูบบุหรี่ ได้แก่ แอสเบสตอส (Asbestos) (ตัวอย่าง เช่น ผู้ที่ทำงานในโรงงานผลิตผ้าเบรกรถยนต์ เป็นต้น) และสารอื่น ๆ นอกจากการสูบบุหรี่ ได้แก่ ก๊าซเรดอน (Radon gas) มลภาวะในอากาศ ควันมลภาวะในสิ่งแวดล้อม การฉายรังสีเพื่อรักษา และโรคปอดบางชนิดก็อาจเพิ่มปัจจัยเสี่ยงต่อมะเร็งปอดได้โดยเฉพาะผู้สูบบุหรี่ร่วมด้วย ดังนั้นจะเห็นได้ว่าปัจจัยการเกิดมะเร็งปอดนอกจากบุหรี่แล้วยังมีปัจจัยหลาย ๆ อย่างร่วมกัน โดยอาการที่นำผู้ป่วยโรคมะเร็งปอดมาพบแพทย์ส่วนมากแล้วผู้ป่วยมักเริ่มด้วยอาการทางการหายใจที่เกิดขึ้นใหม่และมากขึ้นเรื่อย ๆ และผู้ป่วยมาด้วยอาการข้างเคียงที่ไม่จำเพาะต่อโรคมะเร็งเช่น อาการไอ น้ำหนักลด อาการเหนื่อย อาการไอเป็นเลือด เจ็บหน้าอก เสียแหบ อาการเบื่ออาหาร อ่อนเพลีย เป็นต้น การให้การวินิจฉัยในระยะเริ่มแรกจึงมีความสำคัญเพื่อให้ผลการรักษาดีขึ้น (สถาบันมะเร็งแห่งชาติ, 2558)

เหมืองข้อมูล (Data mining) เป็นกระบวนการวิเคราะห์ข้อมูล เพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้นๆ (เพชรรัตน์ ม่วงน้อย และคณะ, 2564) การทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงานหลายประเภท เช่น การพยากรณ์พันธุ์ต้นไม้ การพยากรณ์ผู้ใช้บัตรเครดิตรวมถึงการพยากรณ์ผู้ป่วยเพื่อพยากรณ์การอุบัติของโรคต่าง ๆ เช่น โรคมะเร็งเต้านม โรคเบาหวาน โรคหลอดเลือดและหัวใจและโรคอื่นๆ เป็นต้น (Schuh et al., 2020) ในปัจจุบันมีการประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลมาพยากรณ์การเกิดโรค เช่น โรคมะเร็งเต้านม โรคเบาหวาน โรคไฮเปอร์ไทรอยด์ จากฐานข้อมูลต่างๆ (อุกฤษฎ์ ศรีสุข, 2564) แม้ว่าโรคมะเร็งปอดจะพบได้มากขึ้นในปัจจุบันแต่ส่วนใหญ่ของการวินิจฉัยจะพบเมื่อโรคนั้นอยู่ในระยะที่เป็นมากแล้ว จากแนวคิดดังกล่าว ผู้วิจัยจึงมีแนวคิดในการสร้างและเปรียบเทียบประสิทธิภาพของแบบจำลองที่ใช้สำหรับการพยากรณ์การเกิดโรคมะเร็งปอดด้วยเทคนิคเหมืองข้อมูล เพื่อไปสร้างแบบจำลองการพยากรณ์การเกิดโรคมะเร็งปอดในการคัดกรองผู้ป่วยโรคมะเร็งปอดในประเทศไทย อีกทั้งยังสามารถนำผลการวิเคราะห์ที่ได้ไปพัฒนาเป็นระบบสารสนเทศเพื่อใช้สำหรับการพยากรณ์โอกาสที่จะเป็นมะเร็งปอดเบื้องต้นได้

ระเบียบวิธีวิจัย

ขั้นตอนการดำเนินการวิจัย ประกอบด้วย 5 ขั้นตอน ดังนี้

1. การศึกษาข้อมูล

การวิจัยนี้ได้ใช้ชุดข้อมูลการทำนายโรคมะเร็งปอด (Mysar, 2021) ที่ได้ถูกรวบรวมไว้ในเว็บไซต์ https://www.kaggle.com จำนวนข้อมูลทั้งหมด 310 แถว 16 ตัวแปร โดยอยู่ในรูปแบบไฟล์ csv. เพื่อนำข้อมูลไปวิเคราะห์ ซึ่งมีรายละเอียดดังตารางที่ 1

ตารางที่ 1 ตัวแปรที่ใช้ในการศึกษา

ลำดับ	ตัวแปร	คำอธิบาย	ค่า	ประเภทข้อมูล
1	Gender	เพศ	M = Male, F = Female	Binominal
2	Age	อายุ	ค่าจริง	Integer
3	Smoking	การสูบบุหรี่	1 = ใช่, 2 = ไม่ใช่	Binominal
4	Yellow fingers	ภาวะเล็บเหลือง	1 = ใช่, 2 = ไม่ใช่	Binominal
5	Anxiety	ความวิตกกังวล	1 = ใช่, 2 = ไม่ใช่	Binominal
6	Peer pressure	อิทธิพลจากคนรอบข้าง	1 = ใช่, 2 = ไม่ใช่	Binominal
7	Chronic Disease	โรคเรื้อรัง	1 = ใช่, 2 = ไม่ใช่	Binominal
8	Fatigue	ภาวะอ่อนเพลีย	1 = ใช่, 2 = ไม่ใช่	Binominal
9	Allergy	โรคภูมิแพ้	1 = ใช่, 2 = ไม่ใช่	Binominal
10	Wheezing	หายใจเสียงหวีด	1 = ใช่, 2 = ไม่ใช่	Binominal
11	Alcohol	ดื่มแอลกอฮอล์	1 = ใช่, 2 = ไม่ใช่	Binominal
12	Coughing	ไอมีเสมหะ	1 = ใช่, 2 = ไม่ใช่	Binominal
13	Shortness of Breath	หายใจไม่อิ่ม	1 = ใช่, 2 = ไม่ใช่	Binominal
14	Swallowing Difficulty	ภาวะกลืนลำบาก	1 = ใช่, 2 = ไม่ใช่	Binominal
15	Chest pain	อาการเจ็บหน้าอก	1 = ใช่, 2 = ไม่ใช่	Binominal
16	Lung Cancer	โรคมะเร็งปอด	YES = เป็น, NO =ไม่เป็น	Binominal

2. การเตรียมข้อมูล

งานวิจัยนี้ได้ใช้โปรแกรมสำเร็จรูป RapidMiner เวอร์ชั่น 10.1 ในการเตรียมข้อมูล ซึ่งได้มีการปรับเปลี่ยนข้อมูลให้อยู่ในรูปแบบที่เหมาะสมกับการสร้างแบบจำลองประเภทการจำแนกข้อมูล (Classification) โดยผู้วิจัยได้ทำการเตรียมข้อมูลกับชุดข้อมูลทั้งหมด 2 ขั้นตอนดังนี้

2.1 การคัดเลือกข้อมูล (Data selection)

ผู้วิจัยได้ศึกษาตัวแปรจากชุดข้อมูล ดังแสดงในตารางที่ 1 จำนวน 15 ตัวแปร ประกอบด้วย 1) เพศ 2) อายุ 3) การสูบบุหรี่ 4) ภาวะนิ้วเหลือง 5) ความวิตกกังวล 6) อิทธิพลจากคนรอบข้าง 7) โรคเรื้อรัง 8) ภาวะอ่อนเพลีย 9) โรคภูมิแพ้ 10) หายใจเสียงหวีด 11) ดื่มแอลกอฮอล์ 12) ไอมีเสมหะ 13) หายใจไม่อิ่ม 14) ภาวะกลืนลำบาก และ 15) อาการเจ็บหน้าอก ซึ่งตัวแปรทั้ง 15 ตัวนี้จะทำหน้าที่เป็นตัวแปรอิสระ (Independent variable)

2.2 กำหนดหน้าที่ของตัวแปร

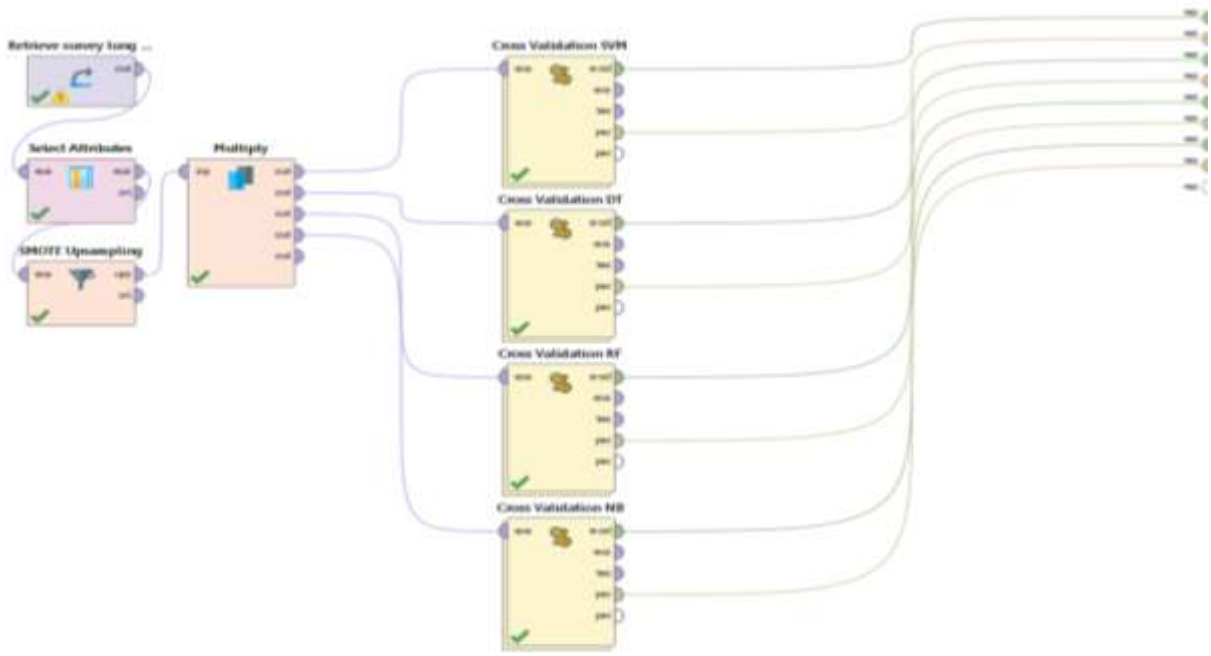
ผู้วิจัยได้กำหนดหน้าที่ให้กับตัวแปรที่ 16 โรคมะเร็งปอด (Lung cancer) กำหนดหน้าที่เป็น “Label” หรือตัวแปรตาม (Dependent variable) เพื่อกำหนดผลลัพธ์ของการพยากรณ์การเป็นโรคมะเร็งปอด

2.3 วิธีสุ่มเพิ่มข้อมูลแบบ SMOTE

เทคนิคการปรับเพิ่มข้อมูลด้วยวิธีสุ่มแบบ SMOTE จัดการความไม่สมดุลของข้อมูล (Imbalanced data) ที่มีปริมาณแตกต่างกัน ข้อมูลในการวิจัยครั้งนี้มีข้อมูลคลาส YES จำนวน 270 (87%) รายการ ส่วนคลาส No จำนวน 39 (13%) รายการ อัตราความไม่สมดุลของข้อมูล (Imbalance ratio) เท่ากับ 6.9 ซึ่งมีผลต่อประสิทธิภาพในการจำแนกข้อมูล วิธี SMOTE เป็นการเพิ่มจำนวนข้อมูลในคลาสที่มีปริมาณน้อย (Minority class) โดยการสุ่มหาข้อมูลขึ้นมา 1 รายการจากกลุ่มคลาส YES หลังจากนั้นพิจารณาค่าข้อมูลเพื่อนบ้านใกล้สุด k ค่า แล้วคำนวณหาระยะห่างระหว่างค่าที่สุ่มกับข้อมูลเพื่อนบ้านใกล้สุดแต่ละค่า เพื่อหาค่าระยะห่างที่น้อยที่สุดระหว่างค่าข้อมูลที่สุ่มกับค่าข้อมูลใกล้เคียงตัวที่มีระยะห่างน้อยสุด (Chawla et al., 2002)

3. การสร้างแบบจำลอง

ในงานวิจัยได้มีการสร้างแบบจำลองประเภทการจำแนกข้อมูล (Classification) จำนวน 4 เทคนิค ประกอบด้วย เทคนิคต้นไม้ตัดสินใจ (Decision tree), เทคนิคป่าสุ่ม (Random forest), เทคนิคนาอีฟเบย์ (Naïve bayes) และเทคนิคซัพพอรตเวกเตอรแมชชีน (Support vector machine)

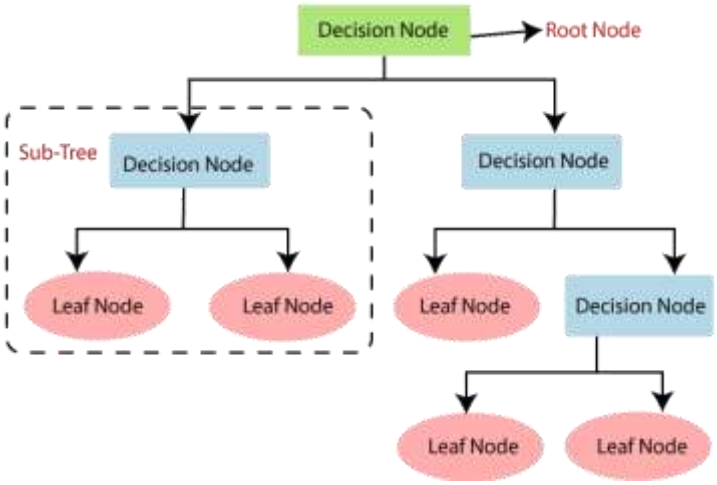


ภาพที่ 1 ขั้นตอนการสร้างแบบจำลองโดยใช้โปรแกรม RapidMiner Studio

(ที่มา : คณะผู้วิจัย, 2566)

3.1 เทคนิคต้นไม้ตัดสินใจ (Decision tree)

เทคนิคต้นไม้ตัดสินใจ เป็นเครื่องมือที่ช่วยให้วิเคราะห์เหตุการณ์ หรือสถานการณ์เพื่อการตัดสินใจได้อย่างเป็นระบบและรวดเร็ว ซึ่งจะแสดงออกมาในรูปแบบของโครงสร้างต้นไม้โดยประกอบไปด้วยกฎในทางรูปแบบ “ถ้า เงื่อนไข แล้ว คำตอบ” โดยโครงสร้างต้นไม้นี้มีคุณลักษณะคล้ายคลึงกับต้นไม้กลับด้าน โดยโหนดแรกสุดซึ่งจะเป็นรากต้นไม้ (Root node) โดยโหนดแสดงคุณลักษณะ (Attribute) กิ่งจะแสดงค่าผลทดสอบและโหนดใบ (Leaf node) โดยคลาสกำหนด แสดงดังภาพที่ 2

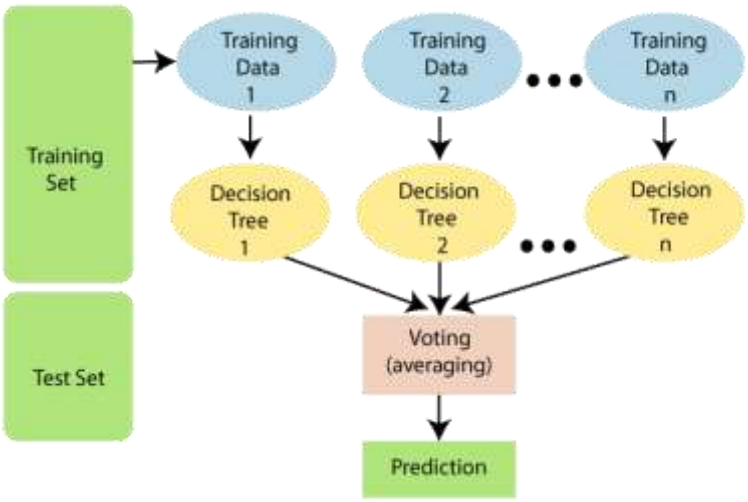


ภาพที่ 2 ตัวอย่างการทำงานของเทคนิคต้นไม้ตัดสินใจ (Decision tree)

(ที่มา : Sonoo Jaiswal, n.d.)

3.2 เทคนิคป่าสุ่ม (Random forest)

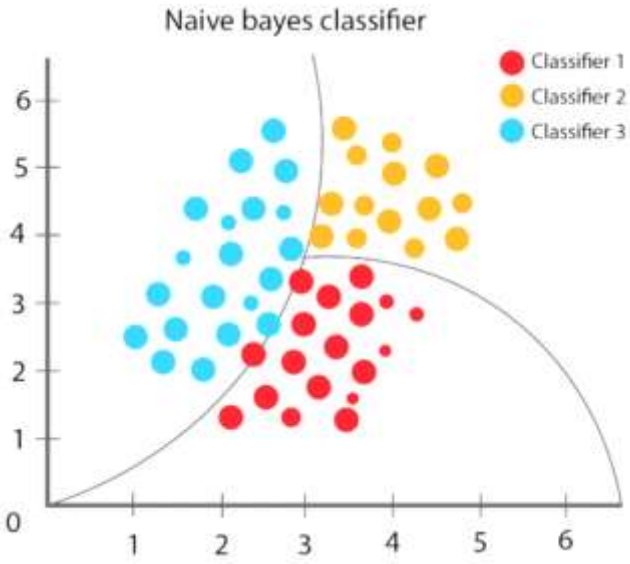
เทคนิคป่าสุ่ม เป็นเทคนิคพัฒนาต่อยอดมาจากเทคนิคต้นไม้ตัดสินใจ โดยจะมีการเพิ่มจำนวนต้นไม้ (Tree) เป็นหลาย ๆ ต้น แต่ละต้นจะได้รับคุณลักษณะ (Feature) และข้อมูล (Data) ที่ไม่เหมือนกันทั้งหมด เพื่อทำให้ได้ต้นไม้ที่มีหลายรูปแบบ และอิสระต่อกันมาก ทำให้ประสิทธิภาพของการทำนายสูงขึ้น องค์ประกอบของเทคนิคการสุ่มป่าไม้จะถูกกำหนดด้วย 3 ส่วนดังนี้ 1) ต้นไม้ทุกต้นจะฝึกสอน (Train) ด้วยวิธีการนำข้อมูลมาย่อยของข้อมูลหลัก 2) เมื่อต้นไม้เริ่มมีขนาดใหญ่มากขึ้นก็จะสามารถค้นหาโหนด (Node) ในแต่ละโหนดที่อยู่ในกิ่งที่ดีมากที่สุดโดยใช้หลักวิธีสุ่ม เลือกคุณลักษณะ N 3) ต้นไม้ทุกต้นจะไม่ทำการทิ้ง แต่จะทำให้ต้นไม้นั้นมีขนาดใหญ่มากขึ้นไปเรื่อย ๆ จนได้คำตอบที่ดีมากที่สุดหลังการสร้างป่า จากนั้นจะทำการให้คะแนน (Vote) โดยต้นไม้ในป่า หากต้นใดได้คะแนนมากสุด ก็จะนำต้นไม้นั้นมาสร้างเป็นตัวแบบสำหรับการพยากรณ์ต่อไป (ศรราม หงส์พรหม และ จันตรี ผลประเสริฐ, 2563) ดังแสดงในภาพที่ 3



ภาพที่ 3 ตัวอย่างการทำงานของเทคนิคต้นไม้ป่าสุ่ม (Random Forest)
(ที่มา : Sonoo Jaiswal, n.d.)

3.3 เทคนิคนาอีฟเบย์ (Naïve bayes)

เทคนคินาอีฟเบย์ เป็นตัวแบบทำนายการจำแนกประเภทข้อมูล โดยวิเคราะห์หาความน่าจะเป็นของสิ่งที่ยังไม่เคยเกิดขึ้น โดยการคาดเดาจากสิ่งที่เคยเกิดขึ้นมาก่อน ความน่าจะเป็นที่เกิดเหตุการณ์หนึ่งก็ต่อเมื่อเหตุการณ์นึงได้เกิดไปแล้ว กล่าวคือเราสนใจจะหาความน่าจะเป็นที่จะเกิดเหตุการณ์ y ถ้ามีเหตุการณ์ x เกิดขึ้นเล้ว โดยมีสมมติฐานว่าปริมาณของความสนใจขึ้นอยู่กับการกระจายความน่าจะเป็น (Probability distribution) เช่น ความน่าจะเป็นคนที่มีงานแล้ว ได้อนุมัติเงินกู้ หรือโอกาสคนที่ปลอดหนี้บ้านจะได้อนุมัติเงินกู้เป็นเท่าใด ยิ่งมีตัวแปรในการพิจารณาจำนวนมาก การพิจารณาความน่าจะเป็นก็จะมากขึ้นตามไปด้วย (จิราภรณ์ เจริญยิ่ง, 2563) ดังแสดงในภาพที่ 4

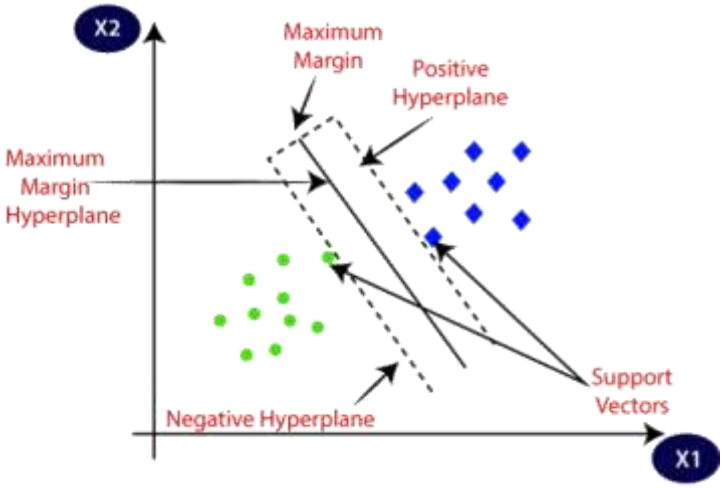


ภาพที่ 4 ตัวอย่างการทำงานของเทคนิคนาอีฟเบย์ (Naïve Bayes)

(ที่มา : Koushiki Dasgupta Chaudhuri, 2023)

3.4 เทคนิคซัพพอรตเวกเตอรแมชชีน (Support vector machine)

เทคนิคซัพพอรตเวกเตอรแมชชีน เป็นอัลกอริธึมในกลุ่มวิธีการเรียนรู้ของเครื่องแบบมีผู้สอน ที่สามารถนำมาช่วยแก้ปัญหาการจำแนกข้อมูลได้ โดยเฉพาะกับปัญหาที่มีขนาดของข้อมูลไม่ใหญ่มาก แต่คุณลักษณะ (Features) ของข้อมูลมีเป็นจำนวนมาก SVM จะถือได้ว่าเป็นอัลกอริธึมที่ทำงานได้ค่อนข้างจะมีประสิทธิภาพมากๆ อัลกอริธึมหนึ่ง หลักการทำงานของ SVM จะอาศัยใช้การสร้างเส้นแบ่ง หรือไฮเปอร์เพลน (Hyperplane) ในการแบ่งแยกคลาสของข้อมูลออกจากกัน จากนั้นจะทำการหาว่าไฮเปอร์เพลนใดเป็นเส้นที่ใช้แยกคลาสของข้อมูลได้ดีที่สุด (Optimal hyperplane) (ไกรศักดิ์ เกษร, 2564) ดังแสดงในภาพที่ 5

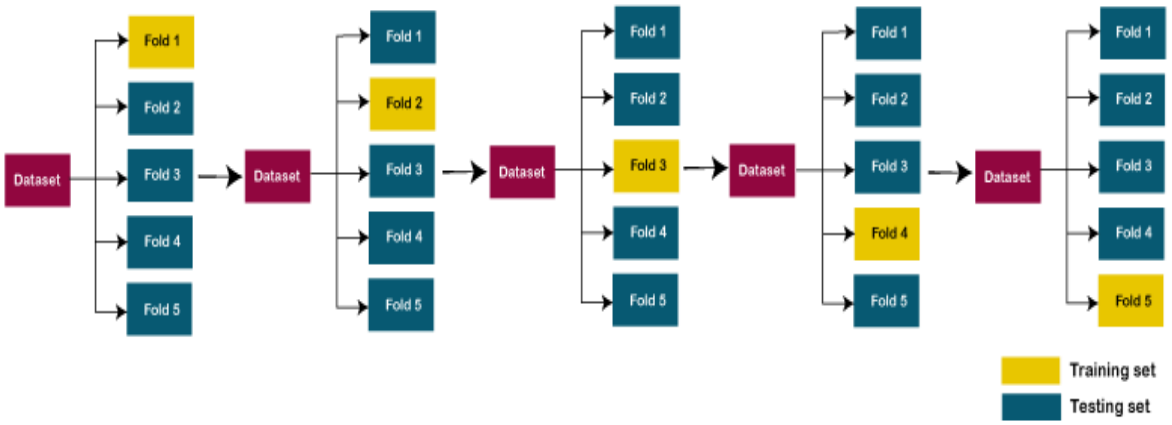


ภาพที่ 5 ตัวอย่างการทำงานของเทคนิคซัพพอรตเวกเตอรแมชชีน (Support vector machine)

(ที่มา : Sonoo Jaiswal, n.d.)

4. การวัดประสิทธิภาพ

การวัดประสิทธิภาพ (Evaluation) เป็นขั้นตอนการประเมินผลลัพธ์ว่ามีความเหมาะสมหรือตรงกับวัตถุประสงค์ที่ต้องการหรือไม่ซึ่งควรนำเสนอผลการวิเคราะห์ในรูปแบบที่ผู้ใช้งานสามารถเข้าใจได้ง่าย วัดประสิทธิภาพของแบบจำลองโดยใช้เทคนิคการวัดประสิทธิภาพแบบ 10-Fold cross validation โดยการแบ่งข้อมูลออกเป็น 10 กลุ่มเท่า ๆ กันโดยในแต่ละรอบการทดสอบจะใช้ข้อมูล 1 ชุด เป็นชุดทดสอบและใช้ชุดที่เหลือเป็นชุดฝึกสอน และในการทดลองครั้งที่สองจะใช้ข้อมูลชุดที่ 2 เป็นชุดข้อมูลทดสอบและให้ข้อมูลชุดที่เหลือเป็นข้อมูลชุดฝึกสอน ทำจนกระทั้งข้อมูลทุกชุดข้อมูลได้ถูกนำมาเป็นชุดข้อมูลทดสอบทั้งหมด ซึ่งจำนวนในการทดสอบมีจำนวนเท่ากับ K ครั้ง โดยผลลัพธ์ที่ได้นั้นจะมาคำนวณหาค่าเฉลี่ยความถูกต้องของการจำแนกข้อมูลในแต่ละรอบ (ธงไชย พ้องเสียง และ จารี ทองคำ, 2565) ดังแสดงในภาพที่ 6



ภาพที่ 6 ตัวอย่างการทดสอบประสิทธิภาพแบบ 10- fold cross validation

(ที่มา : Sonoo Jaiswal, n.d.)

ในการทดสอบประสิทธิภาพแบบ 10-fold ซึ่งจะทำการแบ่งชุดข้อมูลออกเป็น 10 ชุด โดยในแต่ละรอบจะใช้ชุดข้อมูลเพื่อเป็นชุดข้อมูลทดสอบ 1 ชุด และให้ชุดข้อมูลอื่น ๆ เป็นข้อมูลชุดสอน โดยจะทำการทดสอบทั้งหมด 10 รอบ ในการวัดประสิทธิภาพการทำงานในแต่ละขั้นตอนวิธี สามารถวัดได้จากผลของการจำแนกกลุ่มของข้อมูล และสามารถหาค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าประสิทธิภาพโดยรวม (F-measure) ดังนี้

1. ค่าความถูกต้อง (Accuracy) คือ ค่าที่ตัวแบบสามารถพยากรณ์ผู้ป่วยที่จะเกิดโรค และไม่เกิดโรคของข้อมูลทั้งหมดอย่างถูกต้อง ดังสมการ

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

2. ค่าความแม่นยำ (Precision) คือ ความแม่นยำที่ทำนายว่าใช่แล้วถูกต้องมากแค่ไหน ในการทำนายว่าใช่ทั้งหมด ดังสมการ

$$\text{Precision} = \frac{TP}{TP+FP}$$

3. ค่าความระลึก (Recall) คือ จำนวนที่ทำนายถูกกี่ตัวเป็นการวัดความถูกต้องของโมเดล สามารถคำนวณได้จากสมการ

$$\text{Recall} = \frac{TP}{TP+FN}$$

4. ค่าประสิทธิภาพโดยรวม (F-measure) คือ ค่าที่กำเนิดจากการเปรียบเทียบโดย ค่า Precision และค่า Recall ในคลาสเป้าหมาย ดังสมการ

$$\text{F-measure คลาสเป้าหมาย YES} = \frac{(2 * \text{Precision(YES)} * \text{Recall (YES)})}{(\text{Precision(YES)} + \text{Recall (YES)})}$$

$$\text{F-measure คลาสเป้าหมาย NO} = \frac{(2 * \text{Precision(YES)} * \text{Recall (YES)})}{(\text{Precision(YES)} + \text{Recall (YES)})}$$

โดยที่ True Positive (TP) คือ ค่าคลาสของเป้าหมายคือ Yes และแบบพยากรณ์ว่า Yes
False Negatives (FN) คือ ค่าคลาสของเป้าหมายคือ Yes และแบบพยากรณ์ว่า NO
True Negatives (TN) คือ ค่าคลาสของเป้าหมายคือ No และแบบพยากรณ์ว่า NO
False Positive (FP) คือ ค่าคลาสของเป้าหมายคือ No และแบบพยากรณ์ว่า Yes

ผลการวิจัย

การเปรียบเทียบประสิทธิภาพของเทคนิคการทำเหมืองข้อมูลครั้งนี้ทำการวิเคราะห์ตามกระบวนการมาตรฐานในการเหมืองข้อมูล (CRISP-DM) โดยใช้เทคนิคการจำแนกประเภทข้อมูล 4 เทคนิค ประกอบด้วย เทคนิคต้นไม้ตัดสินใจ (Decision tree) เทคนิคป่าสุ่ม (Random forest) เทคนิคนาอีฟเบย์ (Naïve bayes) และเทคนิคซัพพอรตเวกเตอรแมชชีน (Support vector machine) และประเมินประสิทธิของแบบจำลองด้วยค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าประสิทธิภาพโดยรวม (F-measure) และข้อมูลที่ได้เป็นข้อมูลแบบไม่สมดุล (Imbalance data) กล่าวคือมีข้อมูลคลาส YES จำนวน 270 (87%) รายการ ส่วนคลาส NO จำนวน 39 (13%) รายการ อัตราความไม่สมดุลของข้อมูล (Imbalance ratio) เท่ากับ 6.9 ผู้วิจัยจึงได้ทำการสังเคราะห์ข้อมูลเพิ่มด้วยวิธี SMOTE เป็นเทคนิคการปรับเพิ่มข้อมูลด้วยวิธีสุ่มซึ่งเป็น

การเพิ่มจำนวนข้อมูลกลุ่มน้อย เพิ่มขึ้นเป็นจำนวนใกล้เคียงกันจึงทำให้ผลลัพธ์ของประสิทธิภาพดีขึ้นไปด้วย ดังแสดงในตารางที่ 2

ตารางที่ 2 รายการชุดข้อมูลก่อนและหลังการเพิ่มด้วยวิธี SMOTE

คลาส	ข้อมูลตั้งต้น	ข้อมูลที่ผ่าน SMOTE
YES	270 (87%)	270 (50%)
No	39 (13%)	270 (50%)
รวมทั้งสิ้น	**309**	**309**

นำมาสร้างแบบจำลองสำหรับพยากรณ์การเกิดโรคมะเร็งปอดและทำการเปรียบเทียบประสิทธิภาพการจำแนกประเภทข้อมูลทั้ง 4 เทคนิคนี้ด้วยเกณฑ์การวัดประสิทธิภาพทั้ง 4 ค่า ได้แก่ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าประสิทธิภาพโดยรวม (F-measure) ซึ่งผลของการวิเคราะห์ประสิทธิภาพของตัวแบบจำลองสำหรับการพยากรณ์การเกิดโรคมะเร็งปอด พบว่า เทคนิคป่าสุ่มเป็นเทคนิคที่ดีที่สุด ให้ค่าความถูกต้อง 94.63% ค่าความแม่นยำ 92.92% ค่าความระลึก 96.67% และค่าประสิทธิภาพโดยรวม 94.73% รองลงมาคือซัพพอรตเวกเตอรแมชชีน ค่าความถูกต้อง 90.37% ค่าความแม่นยำ 88.98% ค่าความระลึก 92.59% และค่าประสิทธิภาพโดยรวม 90.59% เทคนิคต้นไม้ตัดสินใจ มีค่าความถูกต้อง เท่ากับ 88.89% ค่าความแม่นยำ 90.54% ค่าความระลึก 87.04% และค่าประสิทธิภาพโดยรวม 88.61% และเทคนิคที่ให้ค่าความถูกต้องน้อยที่สุด คือ เทคนิคนาอีฟเบย์ โดยให้ค่าความถูกต้อง เท่ากับ 88.33% ค่าความแม่นยำ 90.88% ค่าความระลึก 85.56% และค่าประสิทธิภาพโดยรวม 87.90% ดังแสดงในตารางที่ 3

ตารางที่ 3 การเปรียบเทียบค่าทดสอบประสิทธิภาพของแบบจำลองสำหรับการพยากรณ์การเกิดโรคมะเร็งปอด

เทคนิค	ประสิทธิภาพของแบบจำลอง			
	ค่าความถูกต้อง (%)	ค่าความแม่นยำ (%)	ค่าความระลึก (%)	ค่าประสิทธิภาพโดยรวม (%)
ต้นไม้ตัดสินใจ	88.89	90.54	87.04	88.61
ป่าสุ่ม	94.63	92.92	96.67	94.73
นาอีฟเบย์	88.33	90.88	85.56	87.90
ซัพพอรตเวกเตอรแมชชีน	90.37	88.98	92.59	90.59

การอภิปรายผล

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อสร้างตัวแบบและเปรียบเทียบประสิทธิภาพของแบบจำลองที่ใช้สำหรับพยากรณ์การเกิดโรคมะเร็งปอด จำนวนข้อมูลทั้งหมด 310 แถว 16 ตัวแปร ทำการวิเคราะห์ตามมาตรฐานในการทำเหมืองข้อมูล (CRISP-DM) แต่เนื่องจากข้อมูลที่ผู้วิจัยได้นำมาวิเคราะห์เป็นข้อมูลแบบไม่สมดุล (Imbalance data) ในการวิจัยครั้งนี้จึงได้ทำการสังเคราะหข้อมูลเพิ่มด้วยวิธี SMOTE (Synthetic minority over-sampling technique) เป็นการสังเคราะหข่อมูลเพิ่มโดยอ่างอิงจากข้อมูลที่มีอยู สามารถสร่างความหลากหลายของข่อมูลได่และสงผลดีต่อการสร่างแบบจำลอง (กิตติภพ แซ่เตีย และ จิรภัทร์ หยกรัตนศักดิ์, 2564) และสอดคล้องเกี่ยวกับงานวิจัยเรื่องการแก้ปัญหาข้อมูลไม่สมดุลของข้อมูลสำหรับการจำแนกผู้ป่วยโรคเบาหวาน ผลการวิจัยพบว่าการแก้ปัญหาข้อมูลไม่สมดุลด้วยวิธีสังเคราะห์ข้อมูลใหม่จะมีประสิทธิภาพดีที่สุดในการจำแนกข้อมูล (วิชญ์วิสิฐ เกษรสิทธิ์ และคณะ, 2561) และทำการวิเคราะห์ข้อมูลด้วยเทคนิคการทำเหมืองข้อมูลทั้งหมด 4 เทคนิคประกอบด้วย เทคนิคต้นไม้ตัดสินใจ เทคนิคนาอีฟเบย์ เทคนิคต้นไม้ป่าสุ่ม และเทคนิคซัพพอรตเวกเตอร์แมชชีน มาใช้ในการสร้างแบบจำลอง พบว่า เทคนิคป่าสุ่มเป็นเทคนิคที่ดีที่สุด ให้ค่าความถูกต้อง 94.63% ค่าความแม่นยำ 92.92% ค่าความระลึก 96.67% และค่าประสิทธิภาพโดยรวม 94.73% ซึ่งมีค่าสูงที่สุดเมื่อเปรียบเทียบกับเทคนิคอื่น ๆ และสอดคล้องเกี่ยวกับงานวิจัยเรื่องการเปรียบเทียบประสิทธิภาพของเทคนิคเหมืองข้อมูลสำหรับพยากรณ์การเกิดโรค มีค่าความถูกต้อง 99.73% (อุกฤษฎ์ ศรีสุข, 2564) และงานวิจัยเรื่องการพยากรณ์โรคเบาหวานด้วยเทคนิคเหมืองข้อมูลให้ค่าความถูกต้อง 99.75% (กฤตกนก ศรีพิมพ์สอ และ กิตติพล วิแสง, 2566) แสดงให้เห็นว่าวิธีเทคนิคป่าสุ่ม (Random forest) มีประสิทธิภาพในการพยากรณ์การเกิดโรคได้ดีที่สุดและมีความเหมาะสมในการนำไปใช้สร้างแบบจำลองความเสี่ยงของการเกิดโรค

บทสรุปและข้อเสนอแนะ

การพัฒนาตัวแบบสำหรับการพยากรณ์การเกิดโรคมะเร็งปอด เพื่อช่วยในคัดกรองผู้ป่วยเบื้องต้นก่อนถึงมือแพทย์และสามารถวางแผนการรักษาจากแพทย์ผู้เชี่ยวชาญด้านโรคมะเร็งต่อไป นอกจากนี้ยังสามารถนำตัวแบบที่มีความแม่นยำนี้ไปพัฒนาเป็นระบบสารสนเทศเพื่อพยากรณ์ผู้ป่วยมะเร็งปอดที่จะเกิดขึ้นในอนาคตได้ ผลการวิเคราะห์ในงานวิจัยนี้สามารถใช้ได้เฉพาะชุดข้อมูลที่ผู้วิจัยนำมาศึกษาเท่านั้น ควรมีการใช้เทคนิคเหมืองข้อมูลอื่น ๆ ที่นอกเหนือจากเทคนิคที่ผู้วิจัยได้ใช้ในงานครั้งนี้ เช่น ตัวแบบการถดถอยลอจิสติก (Logistic regression model)

เอกสารอ้างอิง

กฤตกนก ศรีพิมพ์สอ และ กิตติพล วิแสง. (2566). การพยากรณ์โรคเบาหวานด้วยเทคนิคเหมืองข้อมูล. *วารสารวิชาการการจัดการเทคโนโลยี มหาวิทยาลัยราชภัฏมหาสารคาม, 10*(1), 51-63. https://ph02.tci-thaijo.org/index.php/itm-journal/article/view/248575

กิตติภพ แซ่เตีย และ จิรภัทร์ หยกรัตนศักดิ์. (2564). *การจัดการข้อมูลไม่สมดุลของการทำกลยุทธ์เสนอขายประกันต่อยอดสำหรับผู้ถือบัตรเครดิต.* การประชุมวิชาการระดับชาติ ครั้งที่ 13 มหาวิทยาลัยราชภัฏนครปฐม.

ไกรศักดิ์ เกษร. (2564). *วิทยาศาสตร์ข้อมูล (Data Science).* ภาควิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์มหาวิทยาลัยนเรศวร.

จิราภรณ์ เจริญยิ่ง. (2563). *การพยากรณ์ผลสัมฤทธิ์ทางการเรียนด้วยเทคนิคเหมืองข้อมูลโดยใช้ Rapid Miner.* ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ มหาวิทยาลัยศรีนครินทรวิโรฒ.

จีราพร บวรอารักษ์, อริสา สิทธิโชคพันธ์, สิทธิพงศ์ รักตะเมธากูล และ พรพิศ ยิ้มประยูร. (2562). *การพยากรณ์จำนวนผู้ป่วยโรคมะเร็งปอดสาหรับเพศชายและโรคมะเร็งเต้านมสาหรับเพศหญิงในประเทศไทย.* การประชุมวิชาการระดับชาติ ครั้งที่ 16 มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตกำแพงแสน วันที่ 3-4 ธันวาคม 2562.

ธงไชย พ้องเสียง และ จารี ทองคำ. (2565). *แบบจำลองสำหรับพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูงโดยเทคนิคเหมืองข้อมูล.* ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ มหาวิทยาลัยมหาสารคาม

รักถิ่น เหลาหา. (2553). *การพยากรณ์ความเสี่ยงการเกิดโรคมะเร็งปอดโดยใช้ทฤษฎีของการทำเหมืองข้อมูล.* ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ มหาวิทยาลัยขอนแก่น.

เพชรรัตน์ ม่วงน้อย, จักรพันธ์ พลาผล และ ภรัณยา ปาลวิสุทธ. (2564). ตัวแบบประเมินภาวะความเสี่ยงการเป็นโรคซึมเศร้าของนักศึกษาด้วยเทคนิคเหมืองข้อมูล. *วารสารการประยุกต์ใช้เทคโนโลยีสารสนเทศ, 7*(1), 54-63. https://ph02.tci-thaijo.org/index.php/project-journal/article/view/242196

วิชญ์วิสิฐ เกสรสิทธิ์, วิชิต หล่อจีระชุณห์กุล และ จิราวัลย์ จิตรถเวช. (2561). การแก้ปัญหาข้อมูลไม่สมดุลของข้อมูลสำหรับการจำแนกผู้ป่วยโรคเบาหวาน. *วารสารวิจัย มข. ฉบับบัณฑิตศึกษา, 18*(3), 11-21.

ศรราม หงส์พรหม และ จันตรี ผลประเสริฐ. (2563). *การทำนายระดับความยากจนจากของข้อมูลสำมะโนประชากรด้วยการเรียนรู้ของเครื่อง*. สารนิพนธ์วิทยาศาสตรมหาบัณฑิต (เทคโนโลยีสารสนเทศ), มหาวิทยาลัยศรีนครินทรวิโรฒ

อุกฤษฏ์ ศรีสุข. (2564). การเปรียบเทียบประสิทธิภาพของเทคนิคเหมืองข้อมูลสำหรับอุบัติการของผู้ป่วย. *วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยมหาสารคาม, 40*(2), 157-163. https://li01.tci-thaijo.org/index.php/scimsujournal/article/view/247870

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16*(1), 321-357. https://doi.org/10.1613/jair.995

Koushiki, D. C. (2023). *Building Naive Bayes Classifier from Scratch to Perform Sentiment Analysis*. Retrive from https://www.analyticsvidhya.com/blog/2022/03/building-naive-bayes-classifier-from-scratch-to-perform-sentiment-analysis/

Mysar, A. B. (2021). *Lung Cancer*. Retrive from https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer

Rojanamatin, J., Ukranun, W., Supaattagorn, P., Chaiwiriyabunya, I., Wongsena, M., Chaiwerawattana, A., Laowahutanont, P., Chitapanarux, I., Vatanasapt, P., Greater, S. L., Sangrajrang, S., & Buasom, R. (2021). *Cancer in Thailand volume X 2016-2018*. Bangkok Thailand: National Cancer Institute.

Schuh, G., Prote, J.-P., & Hünnekes, P. (2020). Data mining methods for macro level process planning. *Procedia CIRP, 88*, 48-53. https://doi.org/10.1016/j.procir.2020.05.009

Sonoo Jaiswal. (n.d.). *Decision Tree Classification Algorithm*. Retrive from https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

Sonoo Jaiswal. (n.d.). *Random Forest Algorithm*. Retrive from https://www.javatpoint.com/machine-learning-random-forest-algorithm

Sonoo Jaiswal. (n.d.). *Support Vector Machine Algorithm*. Retrive from https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm