บทความวิจัย (Research Article)

การเปรียบเทียบประสิทธิภาพของตัวแบบสำหรับการพยากรณ์โรคมะเร็งปอด

Comparison of Predictive Models for the Prognosis of Lung Cancer

อรรศฎาวุธ เรืองสวัสดิ์[1] อนุพงศ์ สุขประเสริฐ[1*] ทิวา สินธุกูฏ[1] และ ศิริลักษณ์ ไกยวินิจ[1]
Atsadawut Ruangsawud[1], Anupong Sukprasert[1*], Thiwa Sinthukoot[1], Sirilak Kaiwinit[1]

[1]คณะการบัญชีและการจัดการ มหาวิทยาลัยมหาสารคาม
[1]Mahasarakham Business School, Mahasarakham University
*Corresponding author email: anupong.s@acc.msu.ac.th

วันที่รับบทความ (Received)	วันที่ได้รับบทความฉบับแก้ไข (Revised)	วันที่ตอบรับบทความ (Accepted)
15 มีนาคม 2566	29 เมษายน 2566	24 พฤษภาคม 2566

บทคัดย่อ

งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อศึกษาประสิทธิภาพของเทคนิคเหมืองข้อมูล โดยการสร้าง ตัวแบบเพื่อพยากรณ์โอกาสการเกิดโรคมะเร็งปอด จำนวนข้อมูลทั้งหมด 310 แถว 16 ตัวแปร และนำข้อมูลมาวิเคราะห์ตามมาตรฐานในการทำเหมืองข้อมูล (CRISP-DM) การวิเคราะห์ข้อมูลครั้งนี้ผู้วิจัยทำการแบ่งชุดข้อมูลจำนวน 2 กลุ่ม ด้วยวิธีการ 10-fold Cross Validation สำหรับข้อมูลสอนและข้อมูลทดสอบ และทำการวิเคราะห์ข้อมูลด้วยเทคนิคการทำเหมืองข้อมูลทั้งหมด 4 เทคนิคประกอบด้วย เทคนิคเพื่อนบ้านใกล้ที่สุด (k-Nearest Neighbors) เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคต้นไม้ป่าสุ่ม (Random Forest) และทำการเพิ่มประสิทธิภาพของตัวแบบด้วยเทคนิควิธีการรวมกลุ่ม (Vote Ensemble) วิธีการหาค่าที่เหมาะสมที่สุด (Optimization) ด้วยวิธีด้านวิวัฒนาการ (Evolutionary Algorithms) เพื่อค้นหาค่าที่เหมาะสมที่สุดสำหรับชุดของพารามิเตอร์ในแต่ละตัวแบบและทำการเปรียบเทียบประสิทธิภาพการจำแนกประเภทข้อมูลด้วยค่าความแม่น (Accuracy) ค่าประสิทธิภาพโดยรวม (F-measure) ค่าความไว (Sensitivity) ค่าจำเพาะ (Specificity) และค่าทำนายผลบวก (Positive Predictive Value) โดยใช้โปรแกรม RapidMiner Studio Version 10 ในการสร้างตัวแบบและการวิเคราะห์ข้อมูล ผลการวิจัยพบว่า เทคนิควิธีการรวมกลุ่ม เป็นเทคนิคความเหมาะสมมากที่สุดในการสร้างตัวแบบสำหรับการพยากรณ์โอกาสการเกิดโรคมะเร็งโดยค่าความแม่น มากที่สุดถึง 91.57% อยู่ระดับสูงมาก ค่าประสิทธิภาพ เท่ากับ 60.61% อยู่ระดับปานกลาง ค่าความไว เท่ากับ 51.67% อยู่ระดับปานกลาง และค่าจำเพาะ เท่ากับ 97.42% อยู่ระดับสูงมาก ค่าทำนายผลบวก เท่ากับ 73.08% อยู่ระดับสูง ซึ่งผลการวิจัยนี้สามารถนำไปสร้างเป็นระบบสารสนเทศเพื่อพยากรณ์ผู้ป่วยมะเร็งปอด โดยเป็นการคัดกรองข้อมูลผู้ป่วยเบื้องต้นก่อนถึงมือแพทย์

คำสำคัญ : การเพิ่มประสิทธิภาพ การเปรียบเทียบประสิทธิภาพ เหมืองข้อมูล โรคมะเร็งปอด

Abstract

The purpose of this research is to examine the efficiency of data mining techniques by creating a model to predict the likelihood of lung cancer. In the research procedure, the 310 rows 16 variables data was analyzed under the Cross Standard Process for Data Mining (CRISP-DM) and divided into two groups (teaching and testing data) using the 10-fold cross-validation method. Four data mining techniques consist of k-Nearest Neighbors, Decision Tree, Random Forest, and Vote Ensemble to evaluate algorithms for optimization value for each set of parameters in the model and compare the data classification performance by Accuracy, F-measure, Sensitivity, Specificity and Positive Predictive Value with RapidMiner Studio Version 10. The results show that technique of grouping method is the most suitable technique to build a model for trend prediction of cancer with the highest accuracy of 91.57%, which is very high level. The efficiency value is 60.61%, which is moderate level. The sensitivity is 51.67%, which is moderate level. And the specific value is 97.42%, which is very high level. The positive predictive is 73.08%, which is high level. Finally, the results of this research can be used to build a trend prediction for lung cancer patients. It is a preliminary screening of patient information before reaching the doctor.

Keywords: Optimization, Performance Comparison, Data mining, Lung Cancer

บทนำ

ปัจจุบันโรคมะเร็งปอดเป็นสาเหตุในการเสียชีวิต โรคมะเร็งปอดติดอยู่ใน 5 อันดับของโรคมะเร็งทั้งหมดทั่วโลก [1-2] ในปี 2561 จากการสำรวจสถิติโรคมะเร็งโดยสถาบันมะเร็งแห่งชาติพบว่า มีจำนวนผู้ป่วยโรคมะเร็งปอดเพศชาย 14% เพศหญิง 4.8% ในอนาคตอาจมีผู้ป่วยโรคมะเร็งปอดเพิ่มสูงขึ้นและเสียชีวิตจากโรคมะเร็งปอดเพิ่มสูงขึ้น ผู้ป่วยมักถูกวินิจฉัยว่าเป็นระยะแพร่กระจายสถาบันมะเร็งแห่งชาติรายงานอัตราการเสียชีวิตของผู้ป่วยมะเร็งปอดที่ 5 ปีสูงถึง 78% [3] และในระยะเวลาที่ผ่านมา เทคนิคการรักษาผู้ป่วยมะเร็งปอดมีการพัฒนามากขึ้น โดยเฉพาะการรักษามะเร็งแบบจำเพาะต่อเซลล์ (Targeted therapy) ต่อยีนก่อมะเร็ง Epidermal Growth Factor Receptor (EGFR) ที่มีการกลายพันธุ์ใน exon 19 หรือ exon 21 [4-6] แต่อัตราการเสียชีวิตจากโรคมะเร็งปอดอาจจะสอดคล้องกับปัจจัยอื่น ๆ [7]

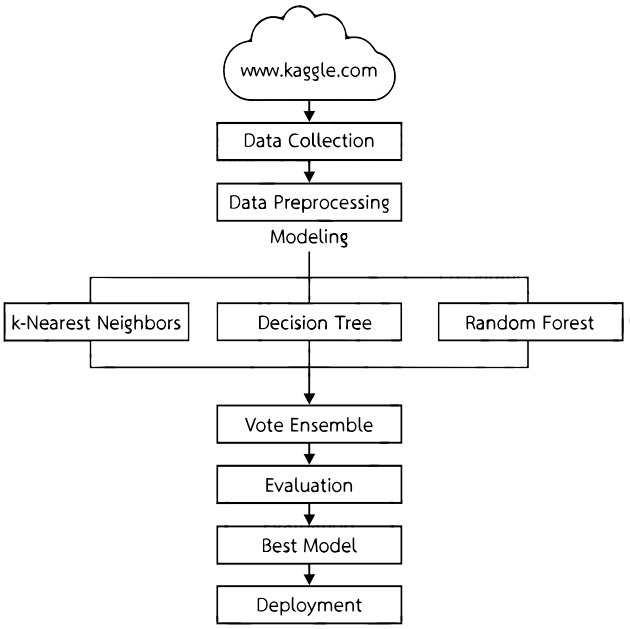
ดังนั้นผู้วิจัยจึงมีความสนใจที่จะพัฒนาตัวแบบสำหรับการคัดกรองผู้ป่วยโรคมะเร็งปอดเบื้องต้นด้วยเทคนิคเหมืองข้อมูล เพื่อไปสร้างตัวแบบพยากรณ์โอกาสการเกิดโรคมะเร็งปอด เพื่อช่วยลดภาระของแพทย์ในการคัดกรองผู้ป่วยโรคมะเร็งปอดในประเทศไทย อีกทั้งยังสามารถนำผลการวิเคราะห์ที่ได้ไปพัฒนาเป็นระบบสารสนเทศเพื่อใช้สำหรับการพยากรณ์โอกาสที่จะเป็นมะเร็งปอดเบื้องต้นได้

วัตถุประสงค์ของการวิจัย

1. เพื่อสร้างตัวแบบสำหรับพยากรณ์โอกาสการเกิดโรคมะเร็งปอด
2. เพื่อเปรียบเทียบตัวแบบที่ใช้สำหรับพยากรณ์โอกาสการเกิดโรคมะเร็งปอด
3. เพื่อเพิ่มประสิทธิภาพตัวแบบที่ใช้สำหรับพยากรณ์โอกาสการเกิดโรคมะเร็งปอด

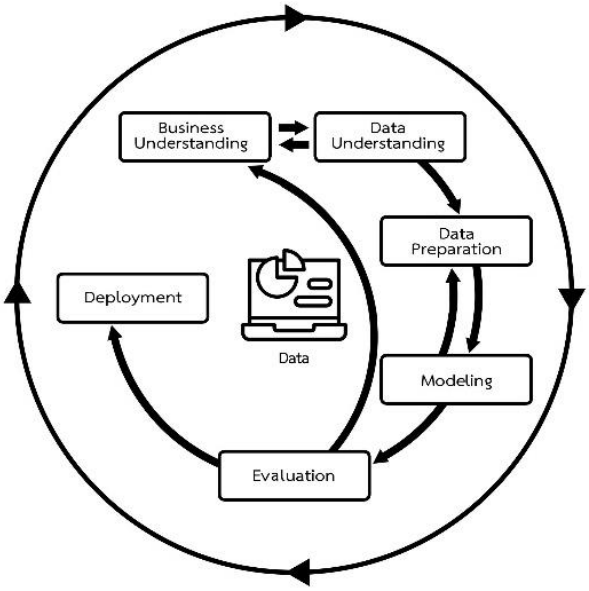
วิธีดำเนินการวิจัย

งานวิจัยนี้มีกรอบแนวคิดสำคัญเพื่อสร้างตัวแบบและเปรียบเทียบประสิทธิภาพของเทคนิคเหมืองข้อมูลที่ใช้สำหรับการสร้างตัวแบบสำหรับการพยากรณ์โอกาสการเกิดโรคมะเร็งปอด เพื่อช่วยคัดกรองผู้ป่วยที่มีโอกาสเกิดโรคมะเร็งปอดเบื้องต้น โดยมีขั้นตอนการดำเนินงานดังนี้



ภาพที่ 1: ขั้นตอนวิธีการดำเนินงานวิจัย

ข้อมูลที่ผู้วิจัยได้นำชุดข้อมูลมาวิเคราะห์ตามมาตรฐานในการทำเหมืองข้อมูล (Cross Standard Process for Data Mining: CRISP-DM) สำหรับการสร้างตัวแบบ การเพิ่มประสิทธิภาพของตัวแบบและเปรียบเทียบเทคนิคเหมืองข้อมูลสำหรับการพยากรณ์โอกาสเกิดโรคมะเร็งปอด ซึ่งประกอบไปด้วย 6 ขั้นตอนดังนี้



ภาพที่ 2: กระบวนการมาตรฐานการทำเหมืองข้อมูล (CRISP-DM)

1. การทำความเข้าใจปัญหา (Business Understanding)

ผู้วิจัยทำการศึกษาข้อมูลที่เกี่ยวข้องกับสาเหตุของการเป็นโรคมะเร็งปอดเกิดจากฝุ่นละอองขนาด PM 2.5 ซึ่งเป็นปัญหาต่อสุขภาพร่างกาย โดยมีแหล่งเกิดจากควันรถยนต์ โรงงานอุตสาหกรรม ผลกระทบที่เกิดขึ้นจากการรับเอาฝุ่นละอองขนาด PM 2.5 เข้าไปในร่างกายปริมาณมากอาจจะส่งผลเสียต่อการแข็งตัวของเลือด การทำงานของเซลล์เยื่อบุหลอดเลือด ทำให้เกิดโรคเรื้อรัง ผลต่อระบบทางเดินหายใจส่วนล่าง เช่น มีอาการไอ อาการเจ็บหน้าอก หายใจไม่สะดวก และอาจส่งผลก่อให้เกิดโรคมะเร็งปอด งานวิจัยนี้จึงต้องการศึกษาตัวแบบที่สามารถนำมาใช้สำหรับการพยากรณ์โอกาสการเกิดโรคมะเร็งปอดของผู้ป่วยด้วยเทคนิคเหมืองข้อมูล ซึ่งผลการศึกษาครั้งนี้สามารถนำไปสร้างเป็นระบบสารสนเทศเพื่อคัดกรองผู้ป่วยมะเร็งปอดเบื้องต้นก่อนถึงมือแพทย์ได้

2. การทำความเข้าใจเกี่ยวกับข้อมูล (Data Understanding)

งานวิจัยนี้นำชุดข้อมูลการทำนายโรคมะเร็งปอดจาก TETSUYA SASAKI [8] จำนวนข้อมูลทั้งหมด 310 แถว 16 ตัวแปร โดยอยู่ในรูปแบบไฟล์ CSV เพื่อนำข้อมูลไปวิเคราะห์ ซึ่งมีรายละเอียดดัง ตารางต่อไปนี้

ตารางที่ 1: ข้อมูลที่ใช้ในงานวิจัย

No	Attribute	Description	Values	Type
1	GENDER	เพศ (M=54%, F=46%)	M = ชาย,F = หญิง	Binominal
2	AGE	อายุ (ค่าจริง)	ค่าจริง	Integer
3	SMOKING	การสูบบุหรี่ (1=44%, 2=56%)	1 = ไม่ใช่, 2 = ใช่	Binominal
4	YELLOW_FINGERS	นิ้วเหลือง (1=44%, 2=56%)	1 = ไม่ใช่, 2 = ใช่	Binominal
5	ANXIETY	ความวิตกกังวล (1=50%, 2=50%)	1 = ไม่ใช่, 2 = ใช่	Binominal
6	PEER_PRESSURE	แรงกดดันจากคนรอบข้าง	1 = ไม่ใช่, 2 = ใช่	Binominal

No	Attribute	Description	Values	Type
		(1=50%, 2=50%)		
7	CHRONIC DISEASE	โรคเรื้อรัง (1=49%, 2=51%)	1 = ไม่ใช่, 2 = ใช่	Binominal
8	FATIGUE	ความเหนื่อยล้า (1=35%, 2=65%)	1 = ไม่ใช่, 2 = ใช่	Binominal
9	ALLERGY	โรคภูมิแพ้ (1=45%, 2=55%)	1 = ไม่ใช่, 2 = ใช่	Binominal
10	WHEEZING	การหายใจถี่ (1=45%, 2=55%)	1 = ไม่ใช่, 2 = ใช่	Binominal
11	ALCOHOL CONSUMING	การบริโภคแอลกอฮอล์ (1=45%, 2=55%)	1 = ไม่ใช่, 2 = ใช่	Binominal
12	COUGHING	อาการไอ (1=42%, 2=58%)	1 = ไม่ใช่, 2 = ใช่	Binominal
13	SHORTNESS OF BREATH	การหายใจไม่สะดวก (1=64%, 2=36%)	1 = ไม่ใช่, 2 = ใช่	Binominal
14	SWALLOWING DIFFICULTY	ความยากลำบากในการกลืน (1=45%, 2=55%)	1 = ไม่ใช่, 2 = ใช่	Binominal
15	CHEST PAIN	อาการเจ็บหน้าอก (1=53%, 2=47%)	1 = ไม่ใช่, 2 = ใช่	Binominal
16	LUNG_CANCER	การเป็นโรคมะเร็งปอด (YES =87%, NO =13%)	YES = เป็น NO =ไม่เป็น	Binominal

3. การเตรียมข้อมูล (Data Preparation)

ขั้นตอนเตรียมข้อมูลเป็นขั้นตอนที่ทำให้เกิดความเชื่อมั่นในคุณภาพข้อมูลที่นำมาใช้ แสดงถึงความเชื่อมั่นของข้อมูลก่อนจะนำไปสร้างตัวแบบการพยากรณ์ในครั้งนี้ โดยผู้วิจัยได้ทำการเตรียมข้อมูลกับชุดข้อมูลทั้งหมด 2 ขั้นตอนดังนี้

3.1 การคัดเลือกข้อมูล (Data Selection)

ผู้วิจัยได้ศึกษาปัจจัยที่ส่งผลทำให้เกิดการเป็นโรคมะเร็งปอด ผู้วิจัยจึงได้ทำการคัดเลือกตัวแปรจากชุดข้อมูลดังแสดงในตารางที่ 1 จำนวน 15 ตัวแปร ประกอบด้วย 1. เพศ 2. อายุ 3. การสูบบุหรี่ 4. นิ้วเหลือง 5. ความวิตกกังวล 6. แรงกดดันจากคนรอบข้าง 7. โรคเรื้อรัง 8. ความเหนื่อยล้า 9. โรคภูมิแพ้ 10. การหายใจถี่ 11. การบริโภคแอลกอฮอล์ 12. อาการไอ 13. การหายใจไม่สะดวก 14. ความยากลำบากในการกลืน 15. อาการเจ็บหน้าอก ซึ่งตัวแปรทั้ง 15 ตัวนี้จะทำหน้าที่เป็นตัวแปรอิสระ (Independent Variable)

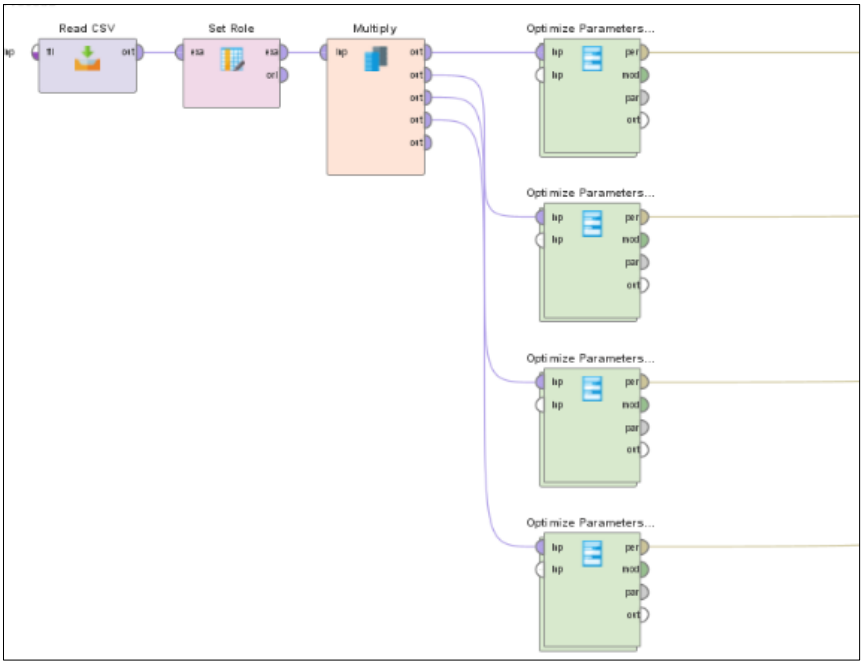
3.2 กำหนดหน้าที่ให้กับตัวแปร

ผู้วิจัยได้กำหนดหน้าที่ให้กับตัวแปรที่ 16 การเป็นโรคมะเร็งปอด (LUNG_CANCER) กำหนดหน้าที่เป็น “Label” หรือตัวแปรตาม (Dependent Variable) เพื่อระบุผลลัพธ์ของการพยากรณ์โอกาสการเป็นโรคมะเร็งปอด

4. การสร้างแบบจำลอง (Modeling)

ขั้นตอนนี้เป็นการสร้างแบบจำลอง ผู้วิจัยได้ใช้โปรแกรม RapidMiner Studio Version 10 [9] มาทำการสร้างตัวแบบสำหรับการพยากรณ์โอกาสการเกิดโรคมะเร็งปอด ด้วยเทคนิคการทำเหมืองข้อมูล 4 เทคนิค ได้แก่ เทคนิคเพื่อนบ้านใกล้ที่สุด (k-Nearest Neighbors) เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคต้นไม้ป่าสุ่ม (Random Forest) และทำการเพิ่มประสิทธิภาพของตัวแบบด้วยเทคนิควิธีการรวมกลุ่ม (Vote Ensemble) วิธีการหาค่าที่เหมาะสมที่สุด (Optimization) ด้วยวิธีด้านวิวัฒนาการ (Evolutionary Algorithms) เพื่อค้นหาค่าที่เหมาะสมที่สุดสำหรับชุดของพารามิเตอร์ในแต่ละตัวแบบ โดยทำการหาค่า k ที่เหมาะสมที่สุดสำหรับเทคนิคเพื่อน

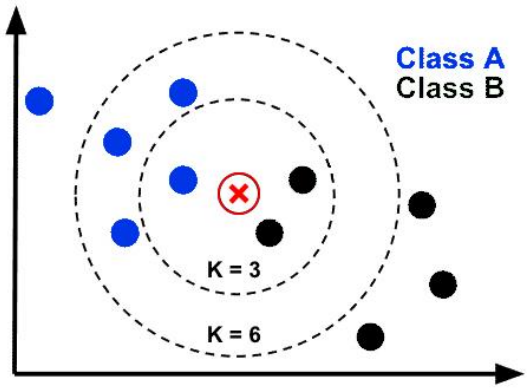
บ้านใกล้ที่สุด และการหาค่าความลึกของต้นไม้แต่ละต้น ที่ดีที่สุดสำหรับเทคนิคต้นไม้ตัดสินใจ และหาจำนวนต้นไม้ที่เหมาะสมสุด สำหรับเทคนิคต้นไม้ป่าสุ่ม ในการสร้างตัวแบบและเพิ่มประสิทธิภาพ แสดงดังภาพ 3



ภาพที่ 3: ขั้นตอนการสร้างตัวแบบและการเพิ่มประสิทธิภาพให้กับตัวแบบโดยใช้โปรแกรม RapidMiner Studio

4.1 เทคนิคเพื่อนบ้านใกล้ที่สุด (k-Nearest Neighbors)

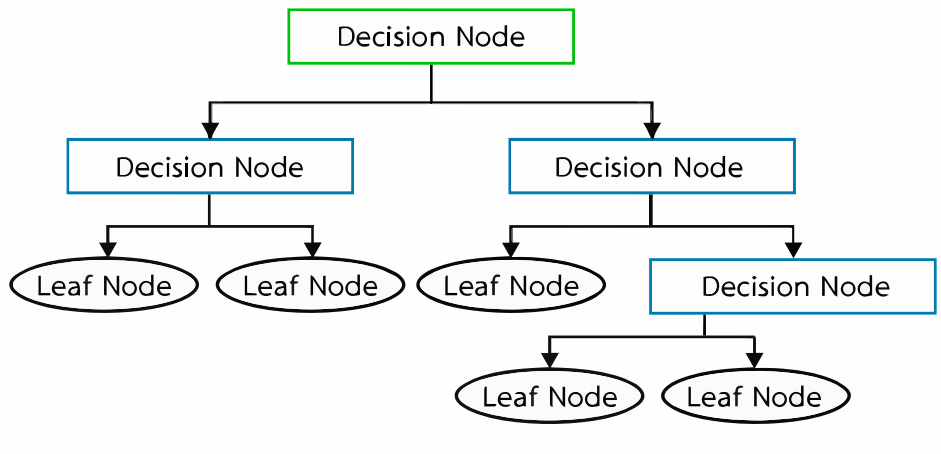
เทคนิคเพื่อนบ้านใกล้ที่สุดเป็นเทคนิคที่มีวิธีไม่ซับซ้อนเข้าใจได้ง่ายที่ใช้ในการจำแนกประเภทข้อมูล โดยจะใช้หลักการเปรียบเทียบความคล้ายคลึงกันของข้อมูลที่เลือกกับข้อมูลอื่น ๆ ว่ามีความคล้ายคลึงหรือใกล้เคียงกับข้อมูลใดมากที่สุด k ตัว จากนั้นจะทำพยากรณ์ว่าคำตอบของข้อมูลที่เลือกนั้นควรเป็นคำตอบเดียวกับข้อมูลที่อยู่ใกล้ที่สุดของ k [9] แสดงดังภาพ 4



ภาพที่ 4: ตัวอย่างการทำงานของเทคนิคเพื่อนบ้านใกล้ที่สุด (k-Nearest Neighbors)

4.2 เทคนิคต้นไม้ตัดสินใจ (Decision Tree)

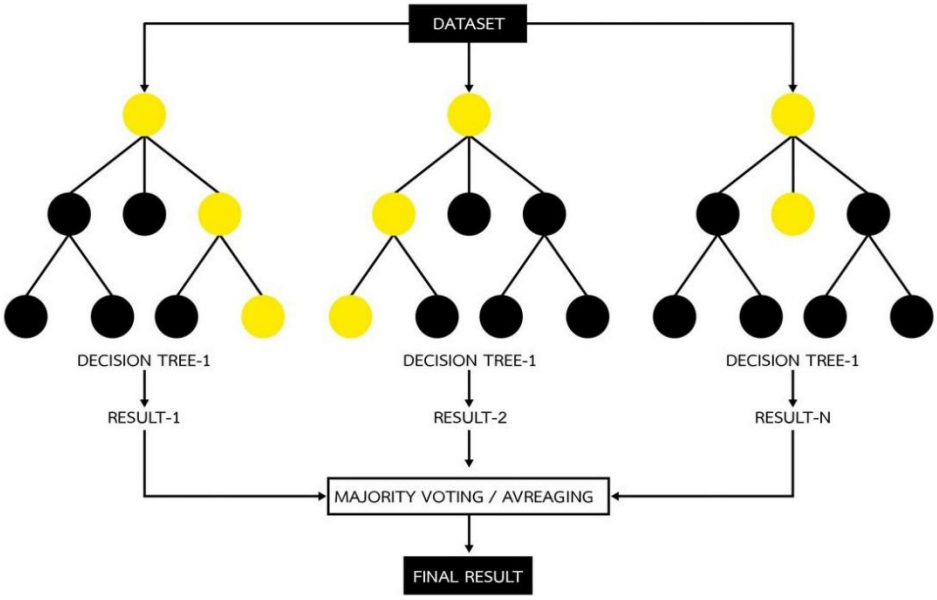
เทคนิคต้นไม้ตัดสินใจ เป็นเครื่องมือที่ช่วยให้วิเคราะห์เหตุการณ์ หรือสถานการณ์เพื่อการตัดสินใจได้อย่างเป็นระบบและรวดเร็ว ซึ่งจะแสดงออกมาในรูปแบบของโครงสร้างต้นไม้โดยประกอบไปด้วยกฎในทางรูปแบบ “ถ้าเงื่อนไข แล้ว คำตอบ” โดยโครงสร้างต้นไม้นี้มีคุณลักษณะคล้ายคลึงกับต้นไม้กลับด้าน โดยโหนดแรกสุดซึ่งจะเป็นรากต้นไม้ (Root node) โดยโหนดแสดงคุณลักษณะ (Attribute) กิ่งจะแสดงค่าผลทดสอบและโหนดใบ (Leaf node) โดยคลาสกำหนด [9] แสดงดังภาพ 5



ภาพที่ 5: ตัวอย่างแผนภาพต้นไม้ของเทคนิคต้นไม้ตัดสินใจ (Decision Tree)

4.3 เทคนิคต้นไม้ป่าสุ่ม (Random Forest)

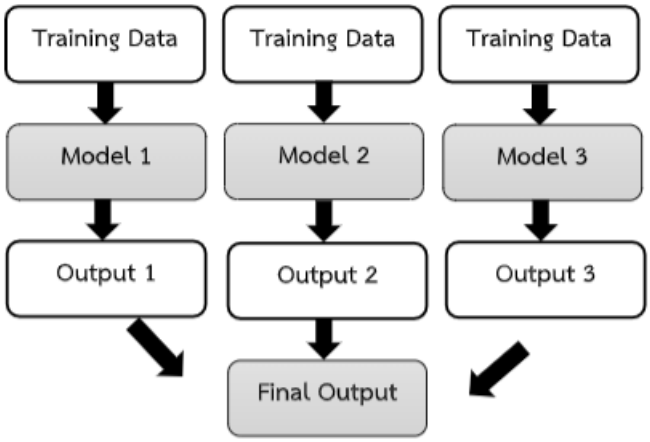
เทคนิคต้นไม้ป่าสุ่ม เป็นเทคนิคพัฒนาต่อยอดมาจาก เทคนิคต้นไม้ตัดสินใจ โดยจะมีการเพิ่มจำนวนต้นไม้ (Tree) เป็นหลาย ๆ ต้น แต่ละต้นจะได้รับคุณลักษณะ (Feature) และข้อมูล (Data) ที่ไม่เหมือนกันทั้งหมด เพื่อทำให้ได้ต้นไม้ที่มีหลายรูปแบบ และอิสระต่อกันมาก ทำให้ประสิทธิภาพของการทำนายสูงขึ้น องค์ประกอบของเทคนิคการสุ่มป่าไม้จะถูกกำหนดด้วย 3 ส่วนดังนี้ 1) ต้นไม้ทุกต้นจะฝึกสอน (Train) ด้วยวิธีการนำข้อมูลมาย่อยของข้อมูลหลัก 2) เมื่อต้นไม้เริ่มมีขนาดใหญ่มากขึ้นก็จะสามารถค้นหาโหนด (Node) ในแต่ละโหนดที่อยู่ในกิ่งที่ดีมากที่สุดโดยใช้หลักวิธีสุ่ม เลือกคุณลักษณะ N 3) ต้นไม้ทุกต้นจะไม่ทำการทิ้ง แต่จะทำให้ต้นไม้นั้นมีขนาดใหญ่มากขึ้นไปเรื่อย ๆ จนได้คำตอบที่ดีมากที่สุดหลังการสร้างป่า จากนั้นจะทำการให้คะแนน (Vote) โดยต้นไม้ในป่า หากต้นใดได้คะแนนมากสุดก็จะนำต้นไม้นั้นมาสร้างเป็นตัวแบบสำหรับการพยากรณ์ต่อไป [10] ดังแสดงดังภาพ 6



ภาพที่ 6: ตัวอย่างการทำงานของเทคนิคต้นไม้ป่าสุ่ม (Random Forest)

4.4 เทคนิควิธีการรวมกลุ่ม (Vote Ensemble)

เทคนิคการรวมกลุ่มเป็นวิธีการหนึ่งของเทคนิคการเรียนรู้ของเครื่อง Machine learning ซึ่งแนวความคิดเกิดจากการรวมกลุ่มของเทคนิคการจำแนกประเภทข้อมูลตั้งแต่ 2 เทคนิคหรือมากกว่า เพื่อสร้างตัวแบบการพยากรณ์ที่เหมาะสมที่สุดมาช่วยในการหาผลลัพธ์ เพื่อแก้ปัญหาเดียวกันและผลที่ได้จะมีความแม่นมากกว่าการใช้โมเดลแบบเดี่ยว ๆ โดยทั่วไปประสิทธิภาพการรวมกลุ่มจะขึ้นอยู่กับความแม่นยำและรูปแบบที่หลากหลายของเทคนิคการจำแนกประเภทข้อมูลที่ใช้ ซึ่งเทคนิควิธีการรวมกลุ่มนั้น สามารถสร้างได้หลากหลายและลดความบกพร่องที่จะเกิดจากความแปรปรวนได้ สำหรับงานวิจัยนี้ผู้วิจัยได้ใช้เทคนิคการจำแนกประเภทข้อมูลจำนวน 3 เทคนิคในการสร้างตัวแบบสำหรับการพยากรณ์โอกาสการเป็นโรคมะเร็งปอดด้วยเทคนิควิธีการรวมกลุ่ม ได้แก่ เทคนิคเพื่อนบ้านใกล้ที่สุด เทคนิคต้นไม้ตัดสินใจ และเทคนิคต้นไม้ป่าสุ่ม และทำการเลือกค่าที่เหมาะสมที่สุดสำหรับเพิ่มประสิทธิภาพสำหรับตัวแบบการจำแนกประเภทข้อมูล เพื่อทำให้เทคนิควิธีการรวมกลุ่มมีประสิทธิภาพสูงขึ้น [9] ดังแสดงดังภาพที่ 7



ภาพที่ 7: รูปแบบการทำงานของเทคนิควิธีการรวมกลุ่ม (Vote Ensemble)

5. การประเมินผล (Evaluation)

ผู้วิจัยทำการแบ่งชุดข้อมูลจำนวน 2 กลุ่ม ด้วยวิธีการ 10-fold cross validation โดยแบ่งข้อมูลออกเป็น 10 กลุ่มเท่า ๆ กัน เพื่อใช้สำหรับเป็นข้อมูลในการสอนและข้อมูลที่ใช้สำหรับการทดสอบตัวแบบ และทำการทดสอบประสิทธิภาพของตัวแบบด้วยค่าความแม่น (Accuracy) ค่าประสิทธิภาพโดยรวม (F-measure) ค่าความไว (Sensitivity) ค่าจำเพาะ (Specificity) และค่าทำนายผลบวก (Positive predictive value) [9] ดังสมการที่ 1-5 ดังนี้

5.1 ค่าความแม่น (Accuracy) คือ ค่าที่ตัวแบบสามารถพยากรณ์ผู้ป่วยที่จะเกิดโรค และไม่เกิดโรคของข้อมูลทั้งหมดอย่างถูกต้อง ดังสมการที่ 1

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

5.2 ค่าประสิทธิภาพโดยรวม (F-measure) คือ ค่าที่กำเนิดจากการเปรียบเทียบโดย ค่า Precision และ ค่า Recall ในคลาสเป้าหมาย ดังสมการที่ 2-3

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

$$\text{F-measure คลาสเป้าหมาย NO} = \frac{(2 * \text{Precision(YES)} * \text{Recall (YES)})}{(\text{Precision(YES)} + \text{Recall (YES)})} \quad (3)$$

5.3 ค่าความไว (Sensitivity) คือ ค่าที่ตัวแบบที่สามารถนำไปพยากรณ์ข้อมูลของผู้ป่วยที่เกิดโรคได้ถูกต้องต่อผู้ป่วยที่เกิดเป็นโรคจริง ดังสมการที่ 4

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

5.4 ค่าจำเพาะ (Specificity) คือ ค่าที่ตัวแบบที่สามารถพยากรณ์ข้อมูลของผู้ป่วยที่ยังไม่เกิดโรคได้ถูกต้องต่อผู้ป่วยที่พยากรณ์สาเหตุเกิดโรค ดังสมการที่ 5

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (5)$$

5.5 ค่าทำนายผลบวก (Positive predictive value) คือ ค่าของตัวแบบที่ทำนายให้ถูกต้อง คำนวณจากจำนวนข้อมูลที่ทำนายถูกในคลาสนั้น จำนวนข้อมูลทั้งหมดที่การทำนายให้ผลลัพธ์เดียวกันในคลาสนั้น ดังสมการที่ 6-7

$$\text{PPV ของคลาสเป้าหมาย YES} = \frac{TP}{TP + FP} \quad (6)$$

$$\text{PPV ของคลาสเป้าหมาย NO} = \frac{TN}{TN + FP} \quad (7)$$

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

สำหรับการวิเคราะห์ข้อมูลครั้งนี้ ผู้วิจัยได้ทำการเพิ่มประสิทธิภาพของตัวแบบการพยากรณ์โอกาสการเกิดโรคมะเร็งปอดด้วยวิธีการหาค่าที่เหมาะสมที่สุด (Optimization) ด้วยวิธีด้านวิวัฒนาการ (Evolutionary Algorithms) เพื่อค้นหาค่าที่เหมาะสมที่สุดสำหรับชุดของพารามิเตอร์ในแต่ละตัวแบบ [11] โดยในเทคนิคเพื่อนบ้านใกล้ที่สุด ได้ทำการหาค่าที่เหมาะสมสำหรับพารามิเตอร์ k เทคนิคต้นไม้ตัดสินใจ ได้ทำการหาค่าที่เหมาะสมสำหรับพารามิเตอร์ค่าความลึกของต้นไม้แต่ละต้นของเทคนิคต้นไม้ตัดสินใจ เทคนิคต้นไม้ป่าสุ่ม ได้ทำการหาค่าที่เหมาะสมสำหรับพารามิเตอร์จำนวนต้นไม้ที่เหมาะสมที่สุดสำหรับเทคนิคต้นไม้ป่าสุ่ม และเทคนิควิธีการรวมกลุ่ม (Vote Ensemble) โดยผู้วิจัยทำการหาค่าที่เหมาะสมสำหรับพารามิเตอร์ของ 3 เทคนิคที่อยู่ภายในเทคนิควิธีการรวมกลุ่มด้วยวิธีการเดียวกันดังที่กล่าวมาข้างต้น

6. การนำไปใช้งาน (Deployment)

เมื่อทำการวิเคราะห์ตามมาตรฐานในการทำเหมืองข้อมูล (CRISP-DM) ผลการวิเคราะห์ข้อมูล พบว่า เทคนิคที่มีความเหมาะสมมากที่สุดในการสร้างตัวแบบการพยากรณ์โรคมะเร็งปอด คือ เทคนิควิธีการรวมกลุ่ม ซึ่งจากผลลัพธ์ของการสร้างแบบจำลองนี้สามารถนำไปใช้สำหรับการพยากรณ์โอกาสการเป็นโรคมะเร็งปอดในประเทศไทย เพื่อช่วยในการคัดกรองผู้ป่วยเบื้องต้นก่อนถึงมือแพทย์ และการวางแผนรักษาเบื้องต้นจากแพทย์ผู้เชี่ยวชาญด้านโรคมะเร็งได้

ผลการวิเคราะห์

งานวิจัยนี้ได้นำข้อมูลปัจจัยที่ส่งผลทำให้เกิดการเป็นโรคมะเร็งปอด 15 ตัวแปร มาทำการสร้างตัวแบบสำหรับการพยากรณ์โรคมะเร็งปอดมาศึกษาตามกระบวนการมาตรฐานในการทำเหมืองข้อมูล (CRISP-DM) เพื่อสร้างตัวแบบการพยากรณ์โอกาสการเกิดโรคมะเร็งปอด โดยผู้วิจัยทำการแบ่งชุดข้อมูลจำนวน 2 ส่วน ส่วนละ 10 กลุ่มเท่า ๆ กัน ด้วยวิธีการ 10-fold Cross validation สำหรับการแบ่งข้อมูลออกเป็นข้อมูลสำหรับการสร้างแบบจำลองและข้อมูลสำหรับการทดสอบตัวแบบ โดยใช้เทคนิคการทำเหมืองข้อมูล 4 เทคนิค ประกอบด้วย เทคนิคเพื่อนบ้านใกล้ที่สุด เทคนิคต้นไม้ตัดสินใจ เทคนิคต้นไม้ป่าสุ่ม และเทคนิควิธีการรวมกลุ่ม จากนั้นทำการทดสอบประสิทธิภาพของการจำแนกประเภทข้อมูลด้วยค่า 5 ค่า ได้แก่ ค่าความแม่น ค่าประสิทธิภาพโดยรวม ค่าความไว ค่าจำเพาะ และค่าทำนายผลบวก เพื่อหาตัวแบบที่มีความหมาะสมสำหรับการพยากรณ์โอกาสการเกิดโรคมะเร็งปอด ซึ่งในการวิเคราะห์ข้อมูลครั้งนี้ ผู้วิจัยได้ทำการเพิ่มประสิทธิภาพของตัวแบบพยากรณ์โอกาสการเกิดโรคมะเร็งปอดด้วยวิธีการหาค่าที่เหมาะสมสำหรับพารามิเตอร์ในแต่ละเทคนิค ซึ่งผลลัพธ์ของการหาค่าที่เหมาะสมที่สุดในแต่ละเทคนิค ได้ค่าดังนี้ โดยเทคนิคเพื่อนบ้านใกล้ที่สุด ค่า k ที่ทำให้ประสิทธิภาพสูงสุดเท่ากับ 2 เทคนิคต้นไม้ตัดสินใจ ค่าความลึกของต้นไม้แต่ละต้นของเทคนิคต้นไม้ตัดสินใจ ที่ทำให้ประสิทธิภาพสูงสุดเท่ากับ 8 เทคนิคต้นไม้ป่าสุ่ม จำนวนต้นไม้ที่เหมาะสมสุดสำหรับเทคนิคต้นไม้ป่าสุ่ม ที่ทำให้ประสิทธิภาพสูงสุดเท่ากับ 6 และเทคนิควิธีการรวมกลุ่มที่มีการรวมกลุ่มของ 3 เทคนิค ได้แก่ เทคนิคเพื่อนบ้านใกล้ที่สุด เทคนิคต้นไม้ตัดสินใจ และเทคนิคต้นไม้ป่าสุ่ม โดยมีการหาค่าที่เหมาะสมที่สุดสำหรับพารามิเตอร์ของ 3 เทคนิค คือ เทคนิคเพื่อนบ้านใกล้ที่สุดค่า k ที่ทำให้ประสิทธิภาพสูงสุดเท่ากับ 10 เทคนิคต้นไม้ตัดสินใจ ค่าความลึกของต้นไม้แต่ละต้นของเทคนิคต้นไม้ตัดสินใจ ที่ทำให้ประสิทธิภาพสูงสุดเท่ากับ 6 และเทคนิคต้นไม้ป่าสุ่ม ค่าจำนวนต้นไม้ที่เหมาะสมสุดสำหรับเทคนิคต้นไม้ป่าสุ่ม ที่ทำให้ประสิทธิภาพสูงสุดเท่ากับ 9 ดังแสดงผลการวิเคราะห์ในตารางที่ 2

ตารางที่ 2: การเปรียบเทียบค่าทดสอบประสิทธิภาพของตัวแบบสำหรับการพยากรณ์โอกาสการเกิดโรคมะเร็งปอด

Classification techniques	Classification performance				
	Accuracy (%)	F-measure (%)	Sensitivity (%)	Specificity (%)	Positive predictive value (%)
k-Nearest Neighbors	90.28%	55.88%	49.17%	96.32%	88.89%
Decision Tree	87.66%	52.50%	54.17%	92.55%	50.33%
Random Forest	90.92%	61.11%	56.67%	95.94%	67.50%
Vote Ensemble*	91.57%	60.61%	51.67%	97.42%	73.08%

* คือ เทคนิคที่มีความเหมาะสมสำหรับนำมาสร้างตัวแบบการพยากรณ์โอกาสเกิดโรคมะเร็งปอดจาก

จากตาราง 2 พบว่า เทคนิควิธีการรวมกลุ่ม เป็นเทคนิคที่ให้ค่าความแม่นสูงสุด โดยมีค่าเท่ากับ 91.57% รองลงมาคือเทคนิคต้นไม้ป่าสุ่ม มีค่าความแม่น เท่ากับ 90.92% และเทคนิคที่ให้ค่าความแม่นน้อยที่สุด คือ เทคนิคต้นไม้ตัดสินใจ โดยให้ค่าความแม่น เท่ากับ 87.66%

ผลสรุปการวิจัย

งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อศึกษาประสิทธิภาพของเทคนิคเหมืองข้อมูล โดยการสร้างตัวแบบเพื่อพยากรณ์โอกาสการเกิดโรคมะเร็งปอด ซึ่งได้ข้อมูลจาก TETSUYA SASAKI [8] จำนวนข้อมูลทั้งหมด 310 แถว 16 ตัวแปร และนำข้อมูลมาวิเคราะห์ตามมาตรฐานในการทำเหมืองข้อมูล (CRISP-DM) การวิเคราะห์ข้อมูลครั้งนี้ผู้วิจัยทำการแบ่งชุดข้อมูลจำนวน 2 กลุ่ม ด้วยวิธีการ 10-fold Cross validation โดยแบ่งข้อมูลออกเป็น 10 กลุ่มเท่า ๆ กัน สำหรับข้อมูลสอนและข้อมูลทดสอบ แต่เนื่องจากข้อมูลที่ผู้วิจัยได้นำมาวิเคราะห์เป็นข้อมูลแบบไม่สมดุล (Imbalance Data) ในปัจจุบันมีหลายวิธีที่ใช้ในการแก้ปัญหาข้อมูลไม่สมดุล เช่น วิธี SMOTE และ Resampling เป็นต้น ซึ่งอยู่นอกขอบเขตของงานวิจัยนี้ โดยผู้วิจัยจะทำการศึกษาเพิ่มเติมต่อไปในอนาคต สำหรับในงานวิจัยนี้ ส่วนข้อมูลที่เป็นข้อมูลไม่สมดุลแบบมือ (Manual) ร่วมกับวิธีการ 10-fold Cross validation โดยแยกข้อมูลที่อยู่ในกลุ่มเป็นโรค กับไม่เป็นโรค มาทำการแบ่ง 10 ส่วนแยกจากกัน แล้วนำข้อมูลมาผสมกันก่อนนำไปใช้เป็นชุดข้อมูลทดสอบป้อนให้กับโมเดลอีกครั้ง และทำการวิเคราะห์ข้อมูลด้วยเทคนิคการทำเหมืองข้อมูลทั้งหมด 4 เทคนิคประกอบด้วย เทคนิคเพื่อนบ้านใกล้ที่สุด เทคนิคต้นไม้ตัดสินใจ เทคนิคต้นไม้ป่าสุ่ม มาใช้ในการสร้างตัวแบบการพยากรณ์ และทำการเพิ่มประสิทธิภาพของตัวแบบด้วยวิธีการหาค่าที่เหมาะสมที่สุด (Optimization) ด้วยวิธีด้านวิวัฒนาการ (Evolutionary Algorithms) เพื่อค้นหาค่าที่เหมาะสมที่สุดสำหรับชุดของพารามิเตอร์ในแต่ละตัวแบบและทำการเปรียบเทียบประสิทธิภาพการจำแนกประเภทข้อมูล โดยนำผลลัพธ์ของงานวิจัยนี้มาสร้างตัวแบบเพื่อพยากรณ์โอกาสการเกิดโรคมะเร็งปอด เพื่อช่วยคัดกรองผู้ป่วยเบื้องต้นก่อนถึงมือแพทย์และวางแผนการรักษาจากแพทย์ผู้เชี่ยวชาญด้านโรคมะเร็งต่อไป นอกจากนี้ยังสามารถนำตัวแบบที่มีความแม่นยำนี้ไปพัฒนาเป็นระบบสารสนเทศเพื่อพยากรณ์ผู้ป่วยมะเร็งปอดที่จะเกิดขึ้นในอนาคตได้

จากผลการวิเคราะห์ข้อมูล พบว่า เทคนิควิธีการรวมกลุ่ม เป็นเทคนิคที่มีความเหมาะสมมากที่สุดในการสร้างตัวแบบสำหรับการพยากรณ์โอกาสการเกิดโรคมะเร็งปอด โดยค่าความแม่นมากที่สุดถึง 91.57% อยู่ระดับสูงมาก ค่าประสิทธิภาพ เท่ากับ 60.61% อยู่ระดับปานกลาง ค่าความไว เท่ากับ 51.67% อยู่ระดับปานกลาง และค่าจำเพาะเท่ากับ 97.42% อยู่ระดับสูงมาก ค่าทำนายผลบวก เท่ากับ 73.08% อยู่ระดับสูง ซึ่งมีค่าสูงที่สุดเมื่อเทียบกับเทคนิคอื่น ๆ และสอดคล้องเกี่ยวกับงานวิจัยของ จิราภา เลาหะวรนันท์ [12] ที่ได้ศึกษาเกี่ยวกับการใช้เทคนิคการทำเหมืองข้อมูลในการจำแนกและคัดเลือกแขนงวิชาสำหรับนักศึกษาคณะเทคโนโลยีสารสนเทศ มีความแม่นสูงถึง 86.67% และงานวิจัยของ พิชญะ พรมลา [13] ที่ได้ศึกษาวิจัยเรื่อง การเปรียบเทียบประสิทธิภาพการวิเคราะห์ความพึงพอใจเกี่ยวกับการจัดการเรียนการสอนด้วยกระบวนการวิเคราะห์ความรู้สึกโดยใช้เทคนิควิธีการรวมกลุ่มเพื่อจำแนกข้อมูลความแม่นสูงถึง 89.06% ซึ่งเปรียบเทียบแล้วมีค่ามากกว่า 91.57%

ข้อเสนอแนะ

ผลการวิเคราะห์ในงานวิจัยนี้สามารถใช้ได้เฉพาะชุดข้อมูลที่ผู้วิจัยนำมาศึกษาเท่านั้น หากมีผู้ที่สนใจศึกษาการสร้างตัวแบบนี้สำหรับการพยากรณ์โรคมะเร็งปอดต่อในอนาคต ควรมีการใช้เทคนิคเหมืองข้อมูลอื่น ๆ ที่นอกเหนือจากเทคนิคที่ผู้วิจัยได้ใช้ในงานครั้งนี้ เพื่อนำไปสร้างตัวแบบในการพยากรณ์ หรือเพิ่มตัวแปรอื่น ๆ ที่ส่งผลต่อการเป็นโรคมะเร็งปอดที่ยังไม่ได้นำมาศึกษาในครั้งนี้

กิตติกรรมประกาศ

งานวิจัยนี้ในได้รับการเอื้อเฟื้อข้อมูลที่เป็นประโยชน์จาก TETSUYA SASAKI เพื่อนำมาใช้ในการวิเคราะห์ข้อมูล และผู้วิจัยขอขอบคุณผู้ช่วยศาสตราจารย์ ดร.นิพจน์พัทธ์ เมืองโคตร ที่ให้ความอนุเคราะห์ ชี้แนะแนวทางให้คำแนะนำปรึกษาตลอดจนปรับปรุงแก้ไขข้อผิดพลาดต่าง ๆ ด้วยความเอาใจใส่อย่างดียิ่ง และคณะการบัญชีและการจัดการมหาวิทยาลัยมหาสารคาม ที่ให้การสนับสนุนในการทำวิจัยครั้งนี้ ขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ ที่นี้

เอกสารอ้างอิง

1 World Health Organization, Cancer fact sheets, lung cancer [อินเตอร์เน็ต]. 2020 [เข้าถึงเมื่อ 7 ธันวาคม 2565]. เข้าถึงจาก : http://gco.iarc.fr/today/fact-sheets-cancers

2 Virani S, Bilheem S, Chansaard W, Chitapanarux I, Daoprasert K, Khuanchana S, et al. National and subnational population-based incidence of cancer in thailand: Assessing cancers with the highest burdens. Cancers (Basel). 2017;9(8).

3 National Cancer Institute. Hospital cancer registration B.E.2561.National Cancer Institute. Department of Medical Services. The Ministry of Public Health. 2019.

4 Non-Small Cell Lung Cancer Collaborative Group. Chemotherapy in non-small cell lung cancer: a meta-analysis using updated data on individual patients from 52 randomised clinical trials. Bmj, 1995, 311.7010: 899-909.

5 Ciuleanu T, Brodowicz T, Zielinski C et al. Maintenance pemetrexed plus best supportive care versus placebo plus best supportive care for non-small-cell lung cancer: a randomised, double-blind, phase 3 study. Lancet 2009; 374: 1432-1440.

6 Schiller JH, Harrington D, Belani CP, Langer C, Sandler A, Krook J, et al. Comparison of four chemotherapy regimens for advanced non-small-cell lung cancer. N Engl J Med. 2002; 346(2): 92–8.

7 วรลักขณ์. วิชพัฒน์, ปัจจัยที่เกี่ยวกับการเสียชีวิต และอัตราการรอดชีวิตในผู้ป่วยโรคมะเร็งปอดชนิดเซลล์ขนาดไม่เล็กระยะแพร่กระจายที่ได้รับการวินิจฉัยและรักษาที่โรงพยาบาลสระบุรี. วารสารกรมการแพทย์, 2563, หน้า 182-192.

8 TETSUYA SASAKI. Survey lung cancer. [อินเตอร์เน็ต]. 2565 [เข้าถึงเมื่อ 7 ธันวาคม 2565]. เข้าถึงจาก : https://www.kaggle.com/code/sasakitetsuya/age-allergy-yellow-fingers-lung-cancer/data?fbclid=IwAR3wFRvO3S1F2V8FxPLmmmcuKKEtgDU8_ssA2wYZ_nE1Ut88f2nDymiV2No

9 อนุพงศ์ สุขประเสริฐ. คู่มือการทำเหมืองข้อมูลด้วยโปรแกรม RapidMiner Studio. พิมพ์ครั้งที่ 4. สาขาวิชาคอมพิวเตอร์ธุรกิจ คณะการบัญชีและการจัดการ มหาวิทยาลัมหาสารคามมหาสารคาม, 2564, หน้า 1-366

10 ศรราม หงส์พรหม และจันตรี ผลประเสริฐ. การทำนายระดับความยากจนจากของข้อมูลสำมะโนประชากรด้วยการเรียนรู้ของเครื่อง. สารนิพนธ์ วท.ม.(เทคโนโลยีสารสนเทศ), มหาวิทยาลัยศรีนครินทรวิโรฒ, 2563, หน้า 1-99.

11 ภรัณยา อำมฤครัตน์ และพยุง มีสัจ, การหาค่าเหมาะสมที่สุดที่มีหลายวัตถุประสงค์ด้วยขั้นตอนวิธีด้านวิวัฒนาการ. วารสารเทคโนโลยีสารสนเทศ, ปีที่ 8, ฉบับที่ 2, กรกฎาคม–ธันวาคม. 2555, หน้า 73-80.

12 จิราภา เลาหะวรนันท์ รชต ลิ้มสุทธิวันภูมิ และบัณฑิต ฐานะโสภณ. การใช้เทคนิคการทำเหมืองข้อมูลในการจำแนกและคัดเลือกแขนงวิชาสำหรับนักศึกษาคณะเทคโนโลยีสารสนเทศ. วารสารเทคโนโลยีสารสนเทศลาดกระบัง, 2558, หน้า 1-9.

13 พิชญะ พรมลา และจรัญ แสนราช. การเปรียบเทียบประสิทธิภาพการวิเคราะห์ความพึงพอใจเกี่ยวกับการจัดการเรียนการสอนด้วยกระบวนการวิเคราะห์ความรู้สึกโดยใช้เทคนิคการรวมกลุ่มเพื่อจำแนกข้อมูล. วารสารวิจัย มข. ฉบับบัณฑิตศึกษา, ปีที่ 20, ฉบับที่ 4, ตุลาคม-ธันวาคม. 2563, หน้า 140-149.