การเพิ่มประสิทธิภาพแบบจำลองเพื่อการพยากรณ์การรักษาโรคความดันโลหิตสูงด้วยการคัดเลือกปัจจัย

Optimization of Models for Hypertension Treatment Prediction with Factor Selection

ธงไชย พ้องเสียง* และ จารี ทองคำ

Tongchai Pongshaing* and Jaree Thongkam

กลุ่มสารสนเทศเชิงประยุกต์ ภาควิชาเทคโนโลยีสารสนเทศ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม

Applied Information Group, Department of Information Technology, Faculty of Information,

Mahasarakham University

*E-mail: tpongshaing@gmail.com

Received: Jun 01, 2022

Revised: Aug 31, 2022

Accepted: Sep 09, 2022

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาปัจจัยที่มีผลต่อการสร้างแบบจำลองเพื่อการพยากรณ์การรักษาโรคความดันโลหิตสูง ในงานวิจัยนี้ได้รวบรวมข้อมูลการรักษาโรคความดันโลหิตสูงจำนวน 2,414 ระเบียนจากฐานข้อมูลโรงพยาบาลทุ่งเขาหลวง อำเภอทุ่งเขาหลวง จังหวัดร้อยเอ็ด จากเดือน มกราคม พ.ศ. 2561 ถึงเดือน ธันวาคม พ.ศ. 2563 ในการเลือกปัจจัยที่เกี่ยวข้องเพื่อเพิ่มประสิทธิภาพของแบบจำลอง วิธี Chi-Square test, Gain Ratio และ Information Gain ได้ถูกนำมาใช้ นอกจากนี้ เทคนิคต้นไม้ตัดสินใจ แบบ C4.5, เทคนิคเคเนียร์เรสเนเบอร์, เทคนิคต้นไม้สุ่ม, เทคนิคเพอร์เซปตรอนหลายชั้น, และเทคนิคลอจิสติกถดถอยแบบต้นไม้ ได้ถูกนำมาใช้ในการสร้างแบบจำลองการพยากรณ์การรักษาโรคความดันโลหิตสูง ในงานวิจัยนี้ผู้วิจัยได้ใช้หลักการ 10-fold cross validation ในการแบ่งข้อมูลออกเป็นชุดเรียนรู้และชุดทดสอบ และใช้ค่าความจำเพาะ ค่าความไว และค่าความถูกต้องของแบบจำลองในการแสดงประสิทธิภาพของแบบจำลอง ผลการทดลองแสดงให้เห็นว่าการคัดเลือกปัจจัยด้วย Gain Ratio เป็นวิธีที่สามารถเพิ่มค่าความจำเพาะและค่าความถูกต้องของแบบจำลองที่สร้างด้วยเทคนิคเคเนียร์เรสเนเบอร์ได้ดีที่สุด โดยสามารถเพิ่มค่าความจำเพาะ และค่าความถูกต้องได้ร้อยละ 15.35 และ 3.72 ตามลำดับ

คำสำคัญ: ความดันโลหิตสูง การพยากรณ์ การคัดเลือกปัจจัย

Abstract

The purpose of this research was to study factors affecting model building to predict hypertension treatment. In this research, 2,414 records of hypertension treatment were collected from Thung Khao Luang Hospital database during January 2018 to December 2020. In order to select relevant factors to increase performance of the prediction models, Chi-Square test, Gain Ratio and Information Gain were applied. In addition, Decision Tree C4.5, K-Nearest Neighbours, Random Tree, Multilayer Perceptron, and Logistic Model Trees techniques were employed to build models for hypertension treatment prediction. In this research, the researchers used 10-fold cross validation to divide data into learning sets and test sets. Specificity, sensitivity and accuracy of the models were used to determine the model performance. The results showed that factor selection by Gain Ratio was the best approach to increase the specificity and accuracy of the model created by K-Nearest Neighbours which the specificity and accuracy were increased by 15.35% and 3.72%, respectively.

Keywords: Hypertension, Prediction, Factor selection

1. บทนำ

โรคความดันโลหิตสูงเป็นปัญหาที่สำคัญของสาธารณะสุขและมีจำนวนผู้ป่วยเพิ่มขึ้นทุกปีทั่วโลกจากข้อมูลรายงานขององค์การอนามัยโลกพบว่ามีผู้ป่วยโรคความดันโลหิตสูงเกือบ 1 พันล้านคนทั่วโลก [1] ส่วนในประเทศไทย โรคความดันโลหิตสูงยังคงเป็นปัญหาสุขภาพที่สำคัญเช่นกัน จากรายงานโรคความดันโลหิตสูงในประชากรอายุ 15 ปีขึ้นไป เพิ่มขึ้นจาก 10 ล้านคน ในปี 2552 เป็น 13 ล้านคน ในปี 2557 จำนวนผู้ป่วยด้วยโรคความดันโลหิตสูงยังมีแนวโน้มเพิ่มขึ้นจาก 4 ล้านคน ในปี 2556 เป็น 6 ล้านคนในปี 2561 [2] การศึกษาเกี่ยวกับการรักษาโรคความดันโลหิตสูงในผู้สูงอายุ พบว่าค่าความดันโลหิตในผู้สูงอายุควรควบคุมให้ไม่เกิน 150/80 มิลลิเมตรปรอท และค่าความดันโลหิตไดแอสโตลิก (Diastolic blood pressure) ที่เกิน 82 มิลลิเมตรปรอท ความดันโลหิตซิสโตลิก (Systolic blood pressure) ที่มากกว่า 148.6 มิลลิเมตรปรอท และน้อยกว่า 126.9 มิลลิเมตรปรอท เสี่ยงที่จะเกิดโรคเรื้อรังตามมา และยังเกิดผลลัพธ์เชิงลบทางคลินิกอีกด้วย [3]

เหมืองข้อมูล (Data Mining) คือกระบวนการวิเคราะห์ข้อมูลจำนวนมาก เพื่อค้นหารูปแบบหรือแบบจำลองที่ซ่อนอยู่ในชุดข้อมูลนั้น ๆ [4], [5], [6] ในปัจจุบันการทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงานหลายประเภท เช่น งานด้านวิทยาศาสตร์ ด้านการทางแพทย์ และการพยากรณ์โรคต่าง ๆ ซึ่งมีนักวิจัยหลายท่านได้นำเอาเทคนิคในเหมืองข้อมูลมาใช้ในการสร้าแบบจำลอง เช่น Khunsuk and Thongkam [7] ได้ทำการเปรียบเทียบประสิทธิภาพของแบบจำลองเพื่อการพยากรณ์การเกิดโรค โรคมะเร็งเต้านม (Breast cancer) โรคเบาหวาน (Diabetes mellitus) โรคไฮเปอร์ไทรอยด์ (Hyperthyroid) จากฐานข้อมูล University of California (UCI) โดยเทคนิคในการคัดเลือกตัวแปรที่ใช้คือ วิธี Chi-Square test หลังจากนั้นจึงวัดประสิทธิภาพของแบบจำลองการพยากรณ์โรคด้วยเทคนิค Decision Tree (DT C4.5) เทคนิค Naïve Bayes (NB) เทคนิค Neural Networks (NN) เทคนิค Random Forest (RF) และเทคนิค Deep Learning (DL) ทดสอบประสิทธิภาพด้วยหลักการ 10-fold cross validation วัดด้วยค่าความถูกต้อง ค่าความไว และค่าความจำเพาะ จากการทดสอบพบว่า เทคนิค Decision Tree (DT C4.5) ให้ค่าความถูกต้องสูงสุดจากชุดข้อมูล ส่วน Aboalnaser and Almohammadi [8] ได้เปรียบเทียบแบบจำลองเพื่อการพยากรณ์การเกิดโรคเบาหวาน ด้วยเทคนิค Naïve Bayes (NB) เทคนิค K-Nearest Neighbours (K-NN) เทคนิค Artificial Neural Network (ANN) เทคนิค Decision Tree (DT) เทคนิค Random Forest (RF) เทคนิค Support Vector Machine (SVM) และเทคนิค Logistic Regression (LR) พบว่า เทคนิค K-Nearest Neighbours (K-NN) มีประสิทธิภาพดีที่สุด ในงานวิจัยของ Alpan and İlgi [9] ได้เสนอการพยากรณ์โรคเบาหวานโดยใช้เทคนิคการทำเหมืองข้อมูล ด้วยเทคนิค Naïve Bayes เทคนิค Decision Tree C4.5 เทคนิค Random Tree เทคนิค Random Forest เทคนิค K-Nearest Neighbours และเทคนิค Support Vector Machine พบว่าเทคนิค Random Tree มีประสิทธิภาพดีที่สุด จะเห็นได้ว่าแต่ละเทคนิคที่นำมาสร้างแบบจำลองที่ผู้วิจัยแต่ละท่านนำมาทดลองมีประสิทธิภาพไม่สูงนัก และมีค่าไม่คงที่เมื่อใช้ข้อมูลที่ไม่ผ่านการคัดเลือกปัจจัย

การคัดเลือกปัจจัยเป็นกระบวนการก่อนการสร้างแบบจำลองที่คัดเลือกตัวแปรที่มีความเกี่ยวข้องกับการพยากรณ์ มาใช้ในการเพิ่มประสิทธิภาพของแบบจำลองในการพยากรณ์โรค เช่น Sujan et al. [10] ได้ทำการคัดเลือกปัจจัยจากข้อมูลผู้ป่วยโรคหลอดเลือดสมองด้วย Chi-Squared และสร้างแบบจำลอง ด้วยเทคนิค Decision Tree C4.5 (C4.5) พบว่าค่าความถูกต้องเพิ่มขึ้นร้อยละ 8.80% ส่วน Sittichat [11] ได้คัดเลือกปัจจัยการศึกษาของนักศึกษาต่อผลการเรียนด้วย Chi-Square และ Gain Ratio สร้างแบบจำลองด้วยเทคนิค Multi-layer perceptron และเทคนิค Decision Tree C4.5 พบว่าค่าความถูกของแบบจำลองที่สร้างด้วยเทคนิค C4.5 เพิ่มขึ้นร้อยละ 13.90 จะเห็นได้ว่าการคัดเลือกปัจจัยแต่ละวิธีสามารถเพิ่มประสิทธิภาพในการสร้างแบบจำลองในเทคนิคที่แตกต่างกัน

ดังนั้นผู้วิจัยจึงสนใจที่จะศึกษาวิธีการคัดเลือกปัจจัยที่เกี่ยวข้อง ด้วยวิธีการ Chi-Square test, Gain Ratio และ Information Gain ในการสร้างแบบจำลองในการพยากรณ์การรักษาโรคความดันโลหิตสูง ผู้วิจัยได้นำเทคนิคเป็นที่นิยมและมีประสิทธิภาพในการสร้างแบบจำลอง คือ เทคนิค C4.5, K-NN, RT, MLP, LMT ในการเปรียบเทียบประสิทธิภาพของวิธีการคัดเลือกปัจจัย ค่าความไว (Sensitivity) ค่าความจำเพาะ (Specificity) และค่าความถูกต้อง (Accuracy) ได้ถูกนำมาใช้ โดยค่าดังกล่าวเป็นค่าที่เหมาะสมกับการพยากรณ์ที่มี 2 คลาส (Binary Class)

2. วิธีการวิจัย

วิธีการดำเนินการวิจัยนี้ได้ใช้ขั้นตอนการทำเหมืองข้อมูลของ [7] ซึ่งมีทั้งหมด 4 ขั้นตอน คือ การเตรียมข้อมูล การคัดเลือกปัจจัย การสร้างแบบจำลอง และการวัดประสิทธิภาพแบบจำลอง

2.1. การเตรียมข้อมูล

การเตรียมข้อมูลในงานวิจัยนี้ได้นำข้อมูลการรักษาโรคความดันโลหิตสูงจากฐานข้อมูลโรงพยาบาลทุ่งเขาหลวง อำเภอทุ่งเขาหลวง จังหวัดร้อยเอ็ด โดยโรงพยาบาลทุ่งเขาหลวง ได้เปิดดำเนินการเมื่อ 1 พฤษภาคม 2556 สังกัดสำนักงานสาธารณสุขจังหวัดร้อยเอ็ด สำนักงานปลัดกระทรวงสาธารณสุข ประเภทโรงพยาบาลระดับ F3 ประชากรในเขตรับผิดชอบประมาณ 23,573 คน และมีผู้ป่วยโรคความดันโลหิตสูงมารับบริการเป็นอันดับต้น ๆ ของโรงพยาบาลจำนวนทั้งหมด 2,639 คน ข้อมูลที่ใช้ในงานวิจัยนี้เป็นข้อมูลที่ผู้ป่วยความดันโลหิตสูงมารับการรักษาในช่วงวันที่ 1 มกราคม 2561 ถึง 31 ธันวาคม 2563 จำนวน 2,414 ระเบียน 13 ปัจจัย โดยปัจจัยบางปัจจัยมีการแปลงเป็นตัวเลข เช่น เพศชาย แปลงเป็น 1, เพศหญิง แปลงเป็น 2, ได้รับยา แปลงเป็น 1, ไม่ได้รับยา แปลงเป็น 0 เป็นต้น ดัง Table 1 และ Table 2

Table 1 แสดงปัจจัยที่มีชนิดข้อมูลเป็นนามบัญญัติ (Nominal) ปัจจัยส่วนใหญ่จะมีเพียง 2 ค่า (Binary values) อาชีพเป็นปัจจัยที่มีจำนวนค่าสูงสุด

Table 2 แสดงปัจจัยที่มีชนิดข้อมูลเป็นตัวเลขเช่น อายุ ความดันขณะหัวใจบีบตัว อัตราชีพจร อัตราการหายใจ ดัชนีมวลการ โดยผู้ป่วยส่วนใหญ่มีอายุประมาณ 68 ปี ส่วนดัชนีมวลกายประมาณ 24.54

เครื่องมือที่นำมาใช้ในการเตรียมข้อมูลได้แก่ โปรแกรม Host Xp โปรแกรม Excel ใช้ในการจัดรูปแบบข้อมูล นำข้อมูลเข้าสู่โปรแกรม โปรแกรม MySQL โปรแกรม PHP เพื่อใช้ในการเตรียมคลาส หรือปัจจัยตาม

Table 1 Nominal data

Variables	Details	Number of value
Sex	เพศ	2
Occ	อาชีพ	32
Marry	สถานะภาพ	6
Edu	ระดับการศึกษา	8
Ht1	ได้รับยาลดความดันโลหิตชนิดที่ 1	2
Ht2	ได้รับยาลดความดันโลหิตชนิดที่ 2	2
Ht3	ได้รับยาลดความดันโลหิตชนิดที่ 3	2
Class	ผลการรักษา (ระดับความดันปกติ,ระดับความดันสูง)	2

Table 2 Numeric data

Variables	Details	Min	Max	Mean	SD
Age	อายุ	11	103	68.23	11.35
Bps	ความดันขณะหัวใจบีบตัว	58	226	134.11	17.32
Pulse	อัตราชีพจร	47	125	81.75	12.01
Rr	อัตราการหายใจ	16	92	20.02	1.48
Bmi	ดัชนีมวลกาย	1.9	83.9	24.54	4.93

2.2. การคัดเลือกปัจจัย

การคัดเลือกปัจจัย เป็นขั้นตอนการเลือกปัจจัยที่มีความเกี่ยวข้องระหว่างปัจจัยสาเหตุ กับปัจจัยตามตามที่มีผลต่อการพยากรณ์การรักษาโรคความดันโลหิตสูง [7] ในงานวิจัยนี้ได้นำโปรแกรม WEKA ในฟังก์ชัน Attribute Selection โดยเลือกปัจจัยที่มีค่าความสำพันธ์มากกว่า 0 ส่วนปัจจัยที่มีความสัมพันธ์เท่ากับ 0 จะถูกตัดออก ปัจจัยที่เหลือจะถูกนำไปสร้างแบบจำลอง และวัดประสิทธิภาพของแบบจำลอง วิธีการที่ถูกนำมาใช้ประกอบด้วย Chi square test, Gain Ratio และ Information Gain

วิธี Chi square test เป็นวิธีการคัดเลือกปัจจัยด้วยการทดสอบความสัมพันธ์ระหว่างปัจจัย (Test of association) ดังสมการที่ 1

$$Chi^2 = \Sigma \frac{(O - E)^2}{E} \qquad (1)$$

โดยที่

Chi^2 คือ ตัวสถิติทดสอบไคสแควร์

O คือ ความถี่ที่ได้จาการสังเกต

E คือ ความถี่ที่คาดหวัง

วิธี Gain Ratio คือวิธีการวัดจำนวนบิตของข้อมูลเพื่อใช้ในการพยาการณ์ ดังสมการที่ 2

$$Gain = Entropy(p) - (\Sigma_{i=0}^{k} \frac{n_i}{n}) Entropy(i)) \qquad (2)$$

โดยที่

$Entropy(p)$ คือ ค่า Entropy ของตัว Root

$(\Sigma_{i=0}^{k} \frac{n_i}{n}) Entropy(i))$ คือ ค่า Entropy ในแต่ละโหนดย่อย

วิธี Information Gain คือ การพิจารณาจากค่าความน่าจะเป็นของแต่ละปัจจัยที่เป็นไปได้แล้ววัดค่าความไร้ระเบียบ (Entropy) เพื่อคัดเลือกปัจจัยที่มีความสำคัญในการจำแนกกลุ่ม ดังสมการที่ 3

$$Entropy(p) = -\Sigma_{i=0}^{c-1} p(j \mid t) \log_2 p(j \mid t) \qquad (3)$$

โดยที่

ΣI คือ ผลรวมของความน่าจะเป็นของค่า j ทีเกิดในคลาส t

$p(j \mid t)$ คือ ค่าความถี่ที่มีความสัมพันธ์ของกลุ่ม j กับโหนด t

ผลการคัดเลือกปัจจัย พบว่าแต่ละวิธีสามารถคัดเลือกได้ 7 ปัจจัยที่มีความสำคัญมากกว่า 0 โดยวิธี Chi-Square test ไม่เลือก Occ วิธี Gain Ratio ไม่เลือก Sex ส่วนวิธี Information Gain ไม่เลือก Edu ผลการคัดเลือกปัจจัยสามารถแสดงได้ดัง Table 3

Table 3 The remaining features after selecting

CST	GR	IG
Bps	Bps	Bps
Ht2	Occ	Occ
Marry	Ht3	Marry
Ht3	Marry	Ht3
Edu	Ht2	Ht2
Ht1	Ht1	Ht1
Sex	Edu	Sex

CST = Chi square test, GR = Gain Ratio and IG = Information Gain

2.3. การสร้างแบบจำลอง

การสร้างแบบจำลองในงานวิจัยนี้ได้นำโปรแกรม WEKA มาสร้างแบบจำลอง 5 เทคนิค ประกอบด้วย เทคนิค Decision tree C4.5 (C4.5) เทคนิค K-Nearest Neighbor (K-NN) เทคนิค Random Tree (RT) เทคนิค Multi-layer perceptron (MLP) และเทคนิค Logistic Model Trees (LMT) ดังนี้

เทคนิค Decision tree C4.5 [12] เป็นวิธีการในการสร้างแบบจำลองโดยแบ่งชุดข้อมูลออกเป็นกลุ่มต่าง ๆ ชุดข้อมูลที่ถูกนำเข้ามา จะเป็นน้ำหนักของแต่ละโหนดเพื่อสร้างตัวแบ่งต่อไป และโหนดสุดท้ายหรือโหนดใบ คือผลลัพธ์ที่ได้

เทคนิค K-Nearest Neighbor [9] เป็นเทคนิคในการสร้างแบบจำลองเพื่อจำแนกชุดข้อมูล จะวิเคราะห์คุณสมบัติใกล้เคียงที่สุดจำนวน k ตัว จะเป็นการคำนวณระยะห่างจุดสองจุดของข้อมูลที่ต้องการจำแนกตามจำนวน k ที่กำหนดไว้ ถ้าข้อมูลไม่คล้ายคลึงกันจะมีระยะที่ห่างออกไป

เทคนิค Random Tree [13] คือ เทคนิคที่ใช้ในการจำแนกหมวดหมู่เช่นเดียวกับ C4.5 โดยมีหลักการสร้างต้นไม้จากการสุ่มหลาย ๆ แบบ ในแต่ละโหนดแล้วเลือกมาประมวลผลโดยไม่ใช้การ Prune จะสุ่มเลือกชุดย่อยของแอทริบิวต์ (Attribute) ก่อนที่จะนำไปใช้

เทคนิค Multi-layer perceptron [14] เป็นเทคนิคในโครงข่ายประสาทเทียมที่มีโครงสร้างแบบหลายชั้น ใช้การคำนวณสูตรทางคณิตศาสตร์โดยการจัดลำดับชั้น โดยที่ข้อมูลต่าง ๆ จะถูกส่งผ่านกลุ่มใยประสาทนำเข้าสู่เซลล์ประสาทเพื่อทำการประมวลผลและส่งผลลัพธ์เพื่อจำแนกข้อมูล

เทคนิค Logistic Model Trees [15] คือการรวมกันของเทคนิค Logistic Regression และเทคนิค Trees ทดสอบคุณลักษณะคือเชื่อมโยงทุกโหนดภายในสำหรับแอทริบิวต์ที่ระบุค่า k และอินสแตนซ์จะถูกจัดเรียงในค่า k และจะขึ้นอยู่กับแอทริบิวต์

2.4. การวัดประสิทธิภาพของแบบจำลอง

การวัดประสิทธิภาพของการพยากรณ์การรักษาโรคความดันโลหิตสูง โดยหลักการใช้ 10-fold cross validation โดยการแบบข้อมูลออกเป็น 10 กลุ่มเท่า ๆ กัน นำ 9 กลุ่มเป็นชุดฝึกสอน และชุดที่เหลือเป็นชุดทดสอบ และทำการเปลี่ยนชุดทดสอบจบครบ 10 รอบ เพื่อหาค่าความ ค่าความไว ค่าความจำเพาะ และค่าความถูกต้อง ของแบบจำลองโดยค่าดังกล่าว เป็นค่าที่ใช้วัดประสิทธิภาพของแบบจำลองประเภทไบล์นารีคลาสได้เป็นอย่างดี [12]

ค่าความไว (Sensitivity) คือ ค่าที่แบบจำลองการรักษาโรคความดันโลหิตสูงสามารถพยากรณ์ผู้ป่วยที่มีระดับความดันปกติได้ถูกต้อง ต่อผู้ป่วยที่มีระดับความดันปกติทั้งหมด ดังสมการที่ 4

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

ค่าความจำเพาะ (Specificity) คือค่าที่แบบจำลองการรักษาโรคความดันโลหิตสูงสามารถพยากรณ์ผู้ป่วยที่มีระดับความดันสูงได้ถูกต้อง ต่อผู้ป่วยที่มีระดับความดันสูงทั้งหมด ดังสมการที่ 5

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

ค่าความถูกต้อง (Accuracy) คือค่าที่แบบจำลองการรักษาโรคความดันโลหิตสูงสามารถพยากรณ์ผู้ป่วยที่มีระดับความดันปกติ และระดับความดันสูงได้ถูกต้อง ต่อผู้ป่วยทั้งหมด ดังสมการที่ 6

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (6)$$

โดยที่

TP คือ จำนวนข้อมูลที่แบบจำลองพยากรณ์การรักษาโรคความดันโลหิตสูงที่ทำให้ผู้ป่วยมีระดับความดันปกติได้อย่างถูกต้อง

TN คือ จำนวนข้อมูลที่แบบจำลองพยากรณ์การรักษาโรคความดันโลหิตสูงที่ทำให้ผู้ป่วยมีระดับความดันสูงได้อย่างถูกต้อง

FP คือ จำนวนข้อมูลที่แบบจำลองพยากรณ์การรักษาโรคความดันโลหิตสูงที่ทำให้ผู้ป่วยมีระดับความดันปกติได้ไม่ถูกต้อง

FN คือ จำนวนข้อมูลที่แบบจำลองพยากรณ์การรักษาโรคความดันโลหิตสูงที่ทำให้ผู้ป่วยมีระดับความดันสูงได้ไม่ถูกต้อง

3. ผลการวิจัยและอภิปรายผลการวิจัย

ในการทดลองครั้งนี้ผู้วิจัยได้ทำการคัดเลือกปัจจัยที่ใช้ในการการพยากรณ์การรักษาโรคความดันโลหิตสูง โดยวิธี Chi-Square test, Gain Ratio และ Information Gain สร้างแบบจำลองด้วยเทคนิค C4.5, K-NN, RF, MLP และ LMT เมื่อใช้ 10-fold cross validation ในการแบ่งข้อมูลเป็นชุดฝึกสอน และหาค่าความไว ค่าความจำเพาะ และค่าความถูกต้องซึ่งเป็นค่าที่แบบจำลองการรักษาโรคความดันโลหิตสูงสามารถพยากรณ์ผู้ป่วยที่มีระดับความดันปกติ และระดับความดันสูงได้ถูกต้อง ต่อผู้ป่วยทั้งหมด

3.1. ค่าความไว

ค่าความไว (Sensitivity) คือค่าที่แบบจำลองการรักษาโรคความดันโลหิตสูงสามารถพยากรณ์ผู้ป่วยที่มีระดับความดันปกติได้ถูกต้อง ต่อผู้ป่วยที่มีระดับความดันปกติทั้งหมด ในการเปรียบเทียบค่าความไว ผู้วิจัยได้เปรียบเทียบแบบจำลองที่สร้างก่อนการคัดเลือกปัจจัยและแบบจำลองที่สร้างหลังการคัดเลือกปัจจัยด้วยวิธี Chi-Square test, Gain Ratio และ Information Gain สร้างแบบจำลองด้วยเทคนิค DT เทคนิค K-NN เทคนิค RT เทคนิค MLP และเทคนิค LMT จากการทดลองสามารถแสดงค่าความไวได้ดัง Table 4

Table 4 แสดงค่าความไวของแบบจำลองการพยากรณ์การรักษาโรคความดันโลหิตสูงก่อนและหลังการคัดเลือกปัจจัยด้วยวิธี Chi-Square test, Gain Ratio และ Information Gain โดยใช้เทคนิค DT เทคนิค K-NN เทคนิค RT เทคนิค MLP และเทคนิค LMT ในการสร้างแบบจำลองการการรักษาโรคความดันโลหิตสูง ผลปรากฏว่า Information Gain สามารถคัดเลือกปัจจัยที่มีความเกี่ยวข้องทำให้ค่าความไวของแบบจำลองที่สร้างด้วยเทคนิค MLP สูงขึ้นที่สุดร้อยละ 2.26 รองลงมา Chi-Square test ช่วยให้แบบจำลองที่สร้างด้วยเทคนิค MLP มีความความไวสูงขึ้นร้อยละ 2.00 อย่างไรก็ตาม Chi-Square test ทำให้แบบจำลองที่สร้างด้วยเทคนิค LMT มีค่าลดลง ส่วน Information Gain ทำให้แบบจำลองที่สร้างด้วยเทคนิค LMT และ C4.5 มีค่าลดลง

3.2. ค่าความจำเพาะ

ค่าความจำเพาะ (Specificity) คือค่าที่แบบจำลองการรักษาโรคความดันโลหิตสูงสามารถพยากรณ์ผู้ป่วยที่มีระดับความดันสูงได้ถูกต้อง ต่อผู้ป่วยที่มีระดับความดันสูงทั้งหมด ในการเปรียบเทียบค่าความจำเพาะผู้วิจัยได้ใช้ข้อมูลก่อนการคัดเลือกปัจจัยและข้อมูลหลังการคัดเลือกปัจจัยด้วยวิธี Chi-Square test, Gain Ratio และ Information Gain เมื่อสร้างแบบจำลองด้วยเทคนิค DT เทคนิค K-NN เทคนิค RT เทคนิค MLP และเทคนิค LMT และทำการทดลองสามารถแสดงค่าความจำเพาะได้ดัง Table 5

Table 5 แสดงค่าความจำเพาะของแบบจำลองการพยากรณ์การรักษาโรคความดันโลหิตสูงก่อนและหลังการคัดเลือกปัจจัยด้วยวิธี Chi-Square test, Gain Ratio, Information Gain โดยใช้เทคนิค เทคนิค DT เทคนิค K-NN เทคนิค RT เทคนิค MLP และเทคนิค LMT ในการสร้างแบบจำลองการการรักษาโรคความดันโลหิตสูง ผลปรากฏว่า Gain Ratio สามารถคัดเลือกปัจจัยที่มีความเกี่ยวข้องทำให้ค่าความจำเพาะของแบบจำลองที่สร้างด้วยเทคนิค K-NN สูงขึ้นที่สุดร้อยละ 15.35 รองลงมา Information Gain ช่วยให้แบบจำที่สร้างด้วยเทคนิค K-NN มีความความจำเพาะสูงขึ้นร้อยละ 12.78 อย่างไรก็ตาม Chi-Square test, Gain Ratio, Information Gain ทำให้แบบจำลองที่สร้างด้วยเทคนิค LMT มีค่าลดลง

3.3. ผลของค่าความถูกต้อง

ค่าความถูกต้อง (Accuracy) คือค่าที่แบบจำลองการรักษาโรคความดันโลหิตสูงสามารถพยากรณ์ผู้ป่วยที่มีระดับความดันปกติ และระดับความดันสูงได้ถูกต้อง ต่อผู้ป่วยทั้งหมด ในการเปรียบเทียบค่าความถูกต้องผู้วิจัยได้ใช้ข้อมูลก่อนการคัดเลือกปัจจัยและข้อมูลหลังการคัดเลือกปัจจัยด้วยวิธี Chi-Square test, Gain Ratio และ Information Gain เมื่อสร้างแบบจำลองด้วยเทคนิค DT เทคนิค K-NN เทคนิค RT เทคนิค MLP และเทคนิค LMT และทำการทดลองสามารถแสดงค่าความถูกต้องได้ดัง Table 6

Table 6 แสดงค่าความถูกต้องของแบบจำลองการพยากรณ์การรักษาโรคความดันโลหิตสูงก่อนและหลังการคัดเลือกปัจจัยด้วยวิธี Chi-Square test, Gain Ratio, Information Gain โดยใช้เทคนิค เทคนิค DT C4.5 เทคนิค K-NN เทคนิค RT เทคนิค MLP และเทคนิค LMT ในการสร้างแบบจำลองการการรักษาโรคความดันโลหิตสูง ผลปรากฏว่า Gain Ratio สามารถคัดเลือกปัจจัยที่มีความเกี่ยวข้องทำให้ค่าความถูกต้องของแบบจำลองที่สร้างด้วยเทคนิค K-NN สูงขึ้นที่สุดร้อยละ 3.72 รองลงมา Information Gain ช่วยให้แบบจำที่สร้างด้วยเทคนิค K-NN มีความความถูกต้องสูงขึ้นร้อยละ 2.98 อย่างไรก็ตาม Gain Ratio ทำให้แบบจำลองที่สร้างด้วยเทคนิค LMT มีค่าลดลง

จากการศึกษาการคัดเลือกปัจจัยในการสร้างแบบจำลองการรักษาโรคความดันโลหิตสูง และเปรียบเทียบประสิทธิภาพของแบบจำลองการรักษาโรคความดันโลหิตสูงด้วยเทคนิคเหมืองข้อมูล ผลการทดลองพบว่าการคัดเลือกปัจจัยด้วยวิธี Gain Ratio สามารถเพิ่มค่าความไว ค่าความจำเพาะ และค่าความถูกต้องของเทคนิค K-NN สูงที่สุด แต่ถึงอย่างไรก็ตามการคัดเลือกปัจจัยด้วยวิธีการ Chi-Square test ร่วมกับเทคนิค DT C4.5 ให้ประสิทธิภาพดีที่สุด เมื่อเทียบกับการคัดเลือกปัจจัยด้วยวิธี Gain Ratio และ Information Gain ร่วมกับเทคนิค K-NN เทคนิค RT เทคนิค MLP และเทคนิค LMT ซึ่งสอดคล้องกับงานวิจัยของ Srisuk and Thongkam [16] ได้ทำการเปรียบเทียบประสิทธิภาพของแบบจำลองของเทคนิคเหมืองข้อมูลสำหรับพยากรณ์การเกิดโรค พบว่า การคัดเลือกปัจจัยด้วยวิธี Chi-Square test ร่วมกับเทคนิค DT C4.5 ให้ค่าความถูกต้องสูงที่สุดเช่นกัน

Table 4 Models' sensitivity

Models	Before feature selection	After feature selection		
		Chi-Square test	Gain Ratio	Information Gain
C4.5	93.70	93.89(↑0.19)	92.30(↓1.40)	92.36(↓1.34)
K-NN	93.06	93.49(↑0.43)	93.27(↑0.21)	93.06(=)
RT	88.97	90.80(↑1.83)	89.56(↑0.59)	89.13(↑0.16)
MLP	91.60	93.60(↑2.00)	91.82(↑0.22)	93.86(↑2.26)
LMT	93.92	93.27(↓0.65)	92.30(↑1.27)	92.84(↓1.08)

Table 5 Models' specificity

Models	Before feature selection	After feature selection		
		Chi-Square test	Gain Ratio	Information Gain
DT C4.5	61.13	62.88(↑1.75)	61.25(↑0.12)	61.83(↑0.70)
K-NN	45.34	56.16(↑10.82)	60.69(↑15.35)	58.12(↑12.78)
RT	48.75	53.91(↑5.61)	51.86(↑3.11)	49.25(↑0.50)
MLP	54.25	60.47(↑6.22)	57.42(↑3.17)	61.62(↑7.37)
LMT	62.95	62.35(↓0.60)	60.61(↓2.34)	61.78(↓1.71)

Table 6 Models' Accuracy

Models	Before feature selection	After feature selection		
		Chi-Square test	Gain Ratio	Information Gain
DT C4.5	79.74	80.82(↑1.08)	80.41(↑0.67)	80.16(↑0.42)
K-NN	76.06	78.38(↑2.32)	79.78(↑3.72)	79.04(↑2.98)
RT	76.55	78.17(↑1.62)	77.59(↑1.04)	76.72(↑0.17)
MLP	78.17	79.58(↑1.41)	79.16(↑0.99)	79.83(↑1.66)
LMT	80.24	80.36(↑0.12)	80.16(↓0.08)	80.36(↑0.12)

4. บทสรุป

จากวัตถุประสงค์เพื่อศึกษาวิธีการคัดเลือกปัจจัยที่มีผลต่อประสิทธิภาพของแบบจำลองเพื่อการพยากรณ์การรักษาโรคความดันโลหิตสูง ด้วยหลักการคัดเลือกตัวแปร วิธี Chi-Square test, Gain Ratio, Information Gain พบว่า การคัดเลือกปัจจัยด้วยวิธี Gain Ratio สามารถเพิ่มค่าความจำเพาะและค่าความถูกต้องของเทคนิค K-NN ถึงร้อยละ 15.35 และร้อยละ 3.72 ตามลำดับ แต่การคัดเลือกปัจจัยด้วยวิธี Gain Ratio ทำให้ค่าความจำเพาะและค่าความถูกต้องของเทคนิค LMT ลดลง และยังทำให้ค่าความไวของเทคนิค C4.5 ลดลง

5. กิตติกรรมประกาศ

ผู้วิจัยขอขอบคุณโรงพยาบาลทุ่งเขาหลวง จังหวัดร้อยเอ็ด ที่ให้ข้อมูลการรักษาผู้ป่วยโรคความดันโลหิตสูงมาศึกษาวิจัยในครั้งนี้ และคณาจารย์ คณะวิทยาการสารสนเทศ ภาควิชาเทคโนโลยีสารสนเทศ มหาวิทยาลัยมหาสารคาม ในการช่วยผลักดันและให้คำปรึกษาคำแนะนำต่าง ๆ ให้งานวิจัยนี้ประสบความสำเร็จ

6. References

[1] Zhou, B. and et al. 2021. Global epidemiology, health burden and effective interventions for elevated blood pressure and hypertension. **Nature Reviews Cardiology.**18: 785-802.

[2] Panmong, N. 2019. **World Hypertension Day Campaign Message Issues.** http://www.thaincd.com. Accessed 19 July 2021. (*in Thai*)

[3] Alsarah, A., Alsara, O. and Bachauwa, G. 2019. Hypertension management in the elderly: What is the optimal target blood pressure? **Heart Views.** 20(1): 11-16.

[4] Somjettana, S. and Thongkam, J. 2021. Performance comparison of data mining techniques for building classification models of the parent toward children who use smart phone. **Journal of Science and Technology, Ubon Ratchathani University.** 23(1): 21-30. (*in Thai*)

[5] Pischarat, K., Ransin, C. and Molpanang, G. 2016. Developing a type 2 diabetes prevention model for community diabetes risk groups Chiang Rai. **Journal of Nursing Science Chulalongkorn University.** 28(3): 132-146. (*in Thai*)

[6] Guttikonda, G., Katamaneni, M. and Pandala, M. 2019. Diabetes data prediction using spark and analysis in Hue over big data. In: **Proceedings of the 3rd International Conference on Computing Methodologies and Communication,** 27-29 March 2019. Erode, India.

[7] Khunsuk, T. and Thongkam, J. 2020. Feature selection method for improving customer reviews classification. **RMUTI JOURNAL Science and Technology.** 13(1): 129-143. (*in Thai*)

[8] Aboalnaser, S. and Almohammadi, H. 2019. Comprehensive study of diabetes miletus prediction using different classification algorithms. In: **Proceedings of the 12th International Conference on Developments in eSystems Engineering,** 7-10 October 2019. Kazan, Russia.

[9] Alpan, K. and İlgi, G. 2020. Classification of diabetes dataset with data mining techniques by using WEKA approach. In: **Proceedings of the 4th International Symposium on Multidisciplinary Studies and Innovative Technologies,** 22-24 October 2020. Istanbul, Turkey.

[10] Sujan, R. and et al. 2020. Chi-squared based feature selection for stroke prediction using AzureML. In: **Proceedings of the 2020 Intermountain Engineering, Technology and Computing,** 2-3 October 2020. Orem, UT, USA.

[11] Sittichat, S. 2017. Study of educational attributes using data mining technique. **Information Technology Journal.** 13(2), 20-28. (*in Thai*)

[12] Khongswat, A. and et al. 2019. The system of stress level prediction using decision tree. **Rattanakosin Journal of Science and Technology.** 1(2): 13-26. (*in Thai*)

[13] Thongkam, J. 2012. Bagging random tree for analyzing breast cancer survival. **KKU Research Journal.** 17(1): 1-13.

[14] Boonta, S. 2019. Forecasting water levels using data mining techniques: A case study at Nong Han Lake, Sakon Nakhon, Thailand. **SNRU Journal of Science and Technology.** 11(2): 64-70.

[15] Landwehr, N., Hall, M. and Frank, E. 2005. Logistic model trees. In: Furnkranz, J. (ed.) **Machine Learning,** Dordrecht, Netherlands: Springer Science+Business Media, Inc.

[16] Srisuk, U. and Thongkam, J. 2021. The efficiency comparison of data mining techniques for patient incidence. **Journal of Science and Technology, Mahasarakham University.** 40(2): 37-43. (*in Thai*)