แบบจำลองสำหรับพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูงโดยเทคนิคเหมืองข้อมูล

วิทยานิพนธ์

ของ

ธงไชย พ้องเสียง

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

ตุลาคม 2565

ลิขสิทธิ์เป็นของมหาวิทยาลัยมหาสารคาม

แบบจำลองสำหรับพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูงโดยเทคนิคเหมืองข้อมูล

วิทยานิพนธ์
ของ
ธงไชย พ้องเสียง

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
ตุลาคม 2565
ลิขสิทธิ์เป็นของมหาวิทยาลัยมหาสารคาม

Prediction Models for Diabetes and High Blood Pressure Treatments using Data Mining Techniques

Tongchai Pongshaing

A Thesis Submitted in Partial Fulfillment of Requirements

for Master of Science (Information Technology)

October 2022

Copyright of Mahasarakham University

คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาวิทยานิพนธ์ของนายธงไชย พ้องเสียง แล้ว เห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยมหาสารคาม

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(ผศ. ดร. ธาดา จันตะคุณ)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผศ. ดร. จารี ทองคำ)

.....กรรมการ
(ผศ. ดร. ฉัตรตระกูล สมบัติธีระ)

.....กรรมการ
(ผศ. ดร. สาธิต แสงประดิษฐ์)

มหาวิทยาลัยอนุมัติให้รับวิทยานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยมหาสารคาม

.....
(ผศ. ศศิธร แก้วมั่น)
คณบดีคณะวิทยาการสารสนเทศ

.....
(รศ. ดร. กริสน์ ชัยมูล)
คณบดีบัณฑิตวิทยาลัย

ชื่อเรื่อง แบบจำลองสำหรับพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูงโดยเทคนิคเหมืองข้อมูล
ผู้วิจัย ธงไชย พ้องเสียง
อาจารย์ที่ปรึกษา ผู้ช่วยศาสตราจารย์ ดร. จารี ทองคำ
ปริญญา วิทยาศาสตรมหาบัณฑิต **สาขาวิชา** เทคโนโลยีสารสนเทศ
มหาวิทยาลัย มหาวิทยาลัยมหาสารคาม **ปีที่พิมพ์** 2565

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์ เพื่อศึกษาปัจจัยการคัดเลือกตัวแปรในการสร้างแบบจำลองเพื่อการพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูง และเปรียบเทียบประสิทธิภาพของแบบจำลองที่สร้างขึ้น งานวิจัยนี้ได้รวบรวมข้อมูลการรักษาโรคความดันโลหิตสูงจากฐานข้อมูลโรงพยาบาลทุ่งเขาหลวง อำเภอทุ่งเขาหลวง จังหวัดร้อยเอ็ด จากเดือน มกราคม พ.ศ.2561 ถึงเดือนธันวาคม พ.ศ.2563 เลือกปัจจัยที่เกี่ยวข้องเพื่อเพิ่มประสิทธิภาพของแบบจำลองด้วย Chi-Square test ได้ถูกนำมาใช้ นอกจากนี้ เทคนิคต้นไม้ตัดสินใจ แบบ C4.5, เทคนิคเคเนียร์เรสเนเบอร์, เทคนิคต้นไม้สุ่ม, เทคนิคเพอร์เซปตรอนหลายชั้น, และเทคนิคลอจิสติกถดถอยแบบต้นไม้ ได้ถูกนำมาใช้ในการสร้างแบบจำลองการพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูง ในการทดลองนี้ผู้วิจัยได้ใช้หลักการ 10-Fold Cross Validation ในการแบ่งชุดข้อมูลออกเป็นชุดเรียนรู้และชุดทดสอบ ในการแสดงประสิทธิภาพของเทคนิคในการคัดเลือกปัจจัยด้วย ค่าความไว ค่าความจำเพาะ และค่าความถูกต้อง จากการทดลองพบว่า ชุดข้อมูลการรักษาโรคเบาหวานที่ทำการคัดเลือกตัวแปรด้วยหลักการ Chi-Square test ด้วยเทคนิค Logistic Model Trees (LMT) ให้ค่าความไว ค่าความจำเพาะ และค่าความถูกต้อง สูงสุดคิดเป็นร้อยละ 95.08, 54.35 และ 70.00 ตามลำดับ ส่วนชุดข้อมูลการรักษาโรคความดันโลหิตสูงที่ทำการคัดเลือกตัวแปรด้วยหลักการ Chi-Square test ด้วยเทคนิค Decision Tree (DT C4.5) ให้ค่าความไว ค่าความจำเพาะ และค่าความถูกต้อง สูงสุดคิดเป็นร้อยละ 93.89, 62.88 และ 80.82 ตามลำดับ

คำสำคัญ : การทำเหมืองข้อมูล, โรคความดันโลหิตสูง, โรคเบาหวาน, ต้นไม้ตัดสินใจ

TITLE Prediction Models for Diabetes and High Blood Pressure Treatments using Data Mining Techniques

AUTHOR Tongchai Pongshaing

ADVISORS Assistant Professor Jaree Thongkam , Ph.D.

DEGREE Master of Science **MAJOR** Information Technology

UNIVERSITY Mahasarakham University **YEAR** 2022

ABSTRACT

The purpose of this research was to study the prognostic modeling of diabetes and hypertension treatment. In this research, data on diabetes and hypertension treatment data were collected from Thung Khao Luang Hospital database. Thung Khao Luang District Roi Et Province From January 2018 to December 2020. To select relevant factors to optimize the model with Chi-Square test, Gain Ratio and Information Gain were used. In addition, the C4.5 decision tree , the K-Nearest Neighbours, the random tree , the multi-layer perceptron , and Logistic Model Trees techniques were employed to build prognostic treatment models. It has been used to model the treatment prognosis for diabetes and hypertension. In this experiment, the researcher used the 10-Fold Cross Validation principle to divide the data set into learning and test sets. To demonstrate the effectiveness of the technique for factor selection by sensitivity, specificity, and accuracy. From the experiment, Diabetes treatment data set using Chi-Square test with Logistic Model Trees (LMT) technique, sensitivity, specificity and the accuracy The highest accounted for 95.08, 54.35 and 70.00 percent, respectively. The data set for the treatment of hypertension with Chi-Square test using Decision Tree (DT C4.5) technique gave sensitivity, specificity and sensitivity. and the accuracy The highest accounted for 93.89, 62.88 and 80.82%, respectively.

Keyword : Data Mining, Hypertension, Diabetes, Decision Tree

พหุน ปณ ฑิโต ชีเว

กิตติกรรมประกาศ

การค้นคว้าวิทยานิพนธ์ฉบับนี้สำเร็จสมบูรณ์ได้ด้วยความกรุณาและความช่วยเหลืออย่างสูงยิ่งจาก ผู้ช่วยศาสตราจารย์ ดร.จารี ทองคำ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.ธาดา จันตะคุณ ประธานกรรมการสอบ ผู้ช่วยศาสตราจารย์ ดร.ฉัตรตระกูล สมบัติธีระ กรรมการสอบ และผู้ช่วยศาสตราจารย์ ดร.สาธิต แสงประดิษฐ์ กรรมการสอบ

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.จารี ทองคำ เป็นอย่างสูงที่ถ่ายทอดวิชาความรู้แก่ศิษย์ด้วยจิตเมตรา กรุณาต่อศิษย์เสมอมา และคอยแนะนำตลอดจนแก้ไขข้อบกพร่องต่างๆด้วยความเอาใจใส่ในทุกขั้นตอน จนวิทยานิพนธ์ฉบับนี้ได้สำเร็จสมบูรณ์ในที่สุด

ขอขอบพระคุณคณาจารย์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม ทุกท่านที่ให้ความรู้คำแนะนำในการศึกษา จนประสบความสำเร็จในการศึกษาครั้งนี้

ขอขอบพระคุณ บิดามารดาญาติพี่น้องตลอดจนถึงผู้ที่คอยสนับสนุนและเป็นกำลังใจ จนทำให้การศึกษาค้นคว้าวิทยานิพนธ์ครั้งนี้สำเร็จไปด้วยดี

และสุดท้ายขอขอบพระคุณ โรงพยาบาลทุ่งเขาหลวง จังหวัดร้อยเอ็ด ที่อนุเคราะห์ข้อมูลในการทำวิจัยในครั้งนี้

ธงไชย พ้องเสียง

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ง

บทคัดย่อภาษาอังกฤษ ..... จ

กิตติกรรมประกาศ ..... ช

สารบัญ ..... ซ

สารบัญภาพ ..... ญ

สารบัญตาราง ..... ฎ

บทที่ 1 บทนำ 1

1.1 หลักการและเหตุผล 1

1.2 วัตถุประสงค์ 4

1.3 ขอบเขตการวิจัย 4

1.4 ประโยชน์ที่คาดว่าจะได้รับ 5

1.5 นิยามศัพท์เฉพาะ 5

บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง 7

2.1 ทฤษฎีที่เกี่ยวข้อง 7

2.1.1 ประวัติโรงพยาบาลทุ่งเขาหลวง 7

2.1.2. โรคเบาหวาน 7

2.1.3 โรคความดันโลหิตสูง 9

2.1.4 การทำเหมืองข้อมูล 10

2.1.5 เทคนิคเหมืองข้อมูลในการวิจัย 11

2.1.6 การวัดประสิทธิภาพ 16

2.2 งานวิจัยที่เกี่ยวข้อง 18

บทที่ 3 วิธีการดำเนินงานวิจัย 20
3.1 การเตรียมข้อมูล 20
3.2 ขั้นตอนก่อนสร้างแบบจำลอง 23
3.3 ขั้นตอนการสร้างแบบจำลอง 24
3.4 ขั้นตอนการวัดประสิทธิภาพ 25
3.5 เครื่องมือที่ใช้ในงานวิจัย 25
บทที่ 4 ผลการวิจัย 26
4.1 ผลการคัดเลือกตัวแปรที่สามารถนำมาพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูง 26
4.2 ผลการเปรียบเทียบประสิทธิภาพของแบบจำลองพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูง 29
บทที่ 5 สรุป อภิปรายและข้อเสนอแนะ 35
5.1 สรุปผลการวิจัย 35
5.2 อภิปรายผล 36
5.3 ข้อเสนอแนะ 36
บรรณานุกรม 37
เอกสารอ้างอิง 38
ภาคผนวก 42
ประวัติผู้เขียน 44

สารบัญภาพ

หน้าที่

ภาพที่ 2.1 ขั้นตอนในการทำเหมืองข้อมูล 10
ภาพที่ 2.2 ตัวอย่างที่ใช้ในการคำนวณหาระยะทาง 13
ภาพที่ 2.3 โครงข่ายเพอร์เซปตรอน 14
ภาพที่ 2.4 โครงข่ายเพอร์เซปตรอนแบบหลายชั้น 15
ภาพที่ 2.5 ตัวอย่างการทดสอบประสิทธิภาพแบบ 10- fold cross validation 16

สารบัญตาราง

หน้าที่

ตารางที่ 3.1 ตัวแปรผู้ป่วยโรคเบาหวาน 21

ตารางที่ 3.2 ตัวแปรผู้ป่วยโรคความดันโลหิตสูง 22

ตารางที่ 4.1 ผลการคัดเลือกตัวแปรผู้ป่วยโรคเบาหวาน 27

ตารางที่ 4.2 ผลการคัดเลือกตัวแปรผู้ป่วยโรคความดันโลหิตสูง 28

ตารางที่ 4.3 ค่าความไวของแบบจำลองผู้ป่วยโรคเบาหวานก่อนและหลังการคัดเลือกตัวแปร 29

ตารางที่ 4.4 ค่าความไวของแบบจำลองผู้ป่วยโรคความดันโลหิตสูงก่อนและหลังการคัดเลือกตัวแปร 30

ตารางที่ 4.5 ค่าความจำเพาะของแบบจำลองผู้ป่วยโรคเบาหวานก่อนและหลังการคัดเลือกตัวแปร 31

ตารางที่ 4.6 ค่าความจำเพาะของแบบจำลองผู้ป่วยโรคความดันโลหิตสูงก่อนและหลังการคัดเลือกตัวแปร 32

ตารางที่ 4.7 ค่าความถูกต้องของแบบจำลองผู้ป่วยโรคเบาหวานก่อนและหลังการคัดเลือกตัวแปร. 33

ตารางที่ 4.8 ค่าความถูกต้องของแบบจำลองผู้ป่วยโรคความดันโลหิตสูงก่อนและหลังการคัดเลือกตัวแปร 34

บทที่ 1
บทนำ

1.1 หลักการและเหตุผล

โรคเบาหวานและโรคความดันโลหิตสูง เป็นโรคที่ประชากรในประเทศไทยเป็นมากเป็นลำดับต้นๆมีผลเสียต่อเศรษฐกิจและสังคม [1] ซึ่งโรคเบาหวานเกิดจากการที่ตับอ่อนไม่สามารถสร้างฮอร์โมนอินซูลินให้เพียงพอต่อความต้องการของร่างกายมนุษย์ได้ ส่วนสาเหตุของโรคเบาหวานมีหลายปัจจัยร่วมกันทั้งปัจจัยทางพันธุกรรม (Genetic factor) และปัจจัยสิ่งแวดล้อม (Environmental factor) เช่น กรรมพันธุ์ พฤติกรรมการกิน ขาดการออกกำลังกาย อาการหลักๆ ที่สื่อว่าคนๆนั้นมีความเสี่ยงที่จะเป็นโรคเบาหวาน ได้แก่ รู้สึกหิวบ่อย กระหายน้ำ ปัสสาวะมีปริมาณมากและบ่อย อีกทั้งก็ยังมีอาการอื่นๆ เช่น เหนื่อย อ่อนเพลีย ร่ายกายซูบผอมลงผิดปกติโดยไม่ทราบสาเหตุ [2] ปัจจุบันความชุกของโรคเบาหวานทั่วโลกเพิ่มสูงขึ้นและมีแนวโน้มเพิ่มขึ้นอย่างต่อเนื่อง [3] โดยความชุกของโรคเบาหวานทั่วโลกในปี 2011 จากข้อมูลของ International Diabetes Federation (IDF) พบความชุกของผู้ป่วยโรคเบาหวานทั่วโลกในปี 2011 ร้อยละ 83 คิดเป็นจำนวน 366 ล้านราย และคาดการณ์ไว้ว่าในปี 2030 จะมีผู้เป็นโรคเบาหวานจำนวน 552 ล้านราย และพบความชุก ของ Impaired Glucose Tolerance (IGT) ในปี 2011 ร้อยละ 6.4 คิดเป็นจำนวน 280 ล้านราย และมีแนวโน้มเพิ่มขึ้นเป็น 398 ล้านรายในปี 2030 โดยทั่วไปแล้วในการวินิจฉัยของแพทย์จะพิจารณาจากค่าน้ำตาลในเลือดของผู้ป่วยหลังจากการให้อดอาหารข้ามคืน (มากกว่า 8 ชั่วโมง) ถ้ามีระดับน้ำตาลมากกว่า หรือ เท่ากับ 126 มิลลิกรัม/เดซิลิตร ถือว่าเป็นผู้ป่วยโรคเบาหวาน หากไม่ได้รับการรักษาอย่างถูกวิธีอาจทำให้เกิดภาวะแทรกซ้อนขึ้นได้ เช่น เบาหวานขึ้นตา ไตวาย กล้ามเนื้อหัวใจขาดเลือด เป็นต้น [4]

ส่วนโรคความดันโลหิตสูง เป็นโรคที่เกิดจากสภาวะการเปลี่ยนแปลงความดันในหลอดเลือดที่สูงขึ้น [5] ซึ่งมีสาเหตุสำคัญมาจากพฤติกรรมสุขภาพที่ไม่เหมาะสม เช่น การรับประทานอาหารที่มีปริมาณโซเดียมสูง การขาดออกกำลังกายรวมถึงการมีพ่อ แม่ หรือญาติพี่น้องสายตรงเป็นโรคความดันโลหิตสูง เป็นต้น ปัจจัยเหล่านี้ล้วนส่งผลให้มีโอกาส ต่อการเกิดโรคนี้มากขึ้น ผู้ป่วยที่มีความดันโลหิตสูงอาจแสดงอาการ แต่ส่วนใหญ่จะมีอาการซึ่งอาจมากหรือน้อย อาการที่พบได้บ่อย คือ ปวดหัว หรือเวียนหัว เหนื่อยง่าย บางครั้งอาจมีอาการของโรคแทรกซ้อนของความดันโลหิตสูง เช่น อัมพาต หรือ มีภาวะหัวใจวาย ปัจจุบันมีผู้ป่วยโรคความดันโลหิตสูงเกือบ 1 พันล้านคนทั่วโลก คาดว่าในปี 2568 ความชุกของผู้ป่วยจะเพิ่มขึ้นเป็น 1.56 พันล้านคน สำหรับสถานการณ์ในประเทศไทย โรคความดันโลหิตสูงยังคงเป็นปัญหาสุขภาพที่สำคัญเช่นกัน เห็นได้จากความชุกของโรคความดันโลหิตสูงในประชากรอายุ 15 ปีขึ้นไปเพิ่มขึ้นจาก 10 ล้านคน ในปี 2552 เป็น 13 ล้านคนในปี 2557 จำนวนผู้ป่วยด้วยโรคความดันโลหิตสูงมีแนวโน้มเพิ่มขึ้นจาก 4 ล้านคนในปี 2556 เป็น 6 ล้านคนในปี 2561 [6] ในการวินิจฉัยของแพทย์ในภาวะบุคคลที่มีความดันเท่ากับหรือมากกว่า 140 มิลลิเมตรปรอท ถือเป็นผู้ป่วยที่มีความดันโลหิตสูง หากปล่อยทิ้งไว้ให้อยู่ในระดับนี้นานๆ โดยที่ไม่ได้รับการรักษาอาจทำให้อวัยวะต่างๆในร่างกายเสื่อมลงได้อย่างรวดเร็ว เช่น หัวใจ สมอง ไต หลอดเลือด และตา [4] จะเห็นว่าโรคเบาหวานและโรคความดันโลหิตสูงยังคงเป็นปัญหาสุขภาพที่สำคัญของประชากรไทย มีรายงานหลายฉบับได้นำสถิติพื้นฐานมาใช้ในการวิเคราะห์แต่ในปัจจุบันเทคนิคเหมืองข้อมูล [7] [8] ได้ถูกนำมาใช้ในการวิเคราะห์ข้อมูลทางการแพทย์เพิ่มมากขึ้น

เหมืองข้อมูล (Data Mining) คือกระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลจำนวนมาก [8] ในปัจจุบันการทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงานหลายประเภท เช่นการพยากรณ์ด้านธุรกิจที่ช่วยในการตัดสินใจของผู้บริหาร ด้านวิทยาศาสตร์และการทางแพทย์ การพยากรณ์โรคต่างๆ มีนักวิจัยหลายท่านที่ได้พยากรณ์โรคเบาหวาน เช่น อุกฤษฎ์ ศรีสุข และจารี ทองคำ [9] ได้ทำการนำเสนอการสร้างแบบจำลองเพื่อการพยากรณ์การเกิดโรค โรคมะเร็งเต้านม โรคเบาหวาน โรคไฮเปอร์ไทรอยด์ จากฐานข้อมูล University of California (UCI) ด้วยเทคนิค Decision Tree (DT C4.5) เทคนิค Naïve Bayes (NB) เทคนิค (Neural Networks: NN) เทคนิค (Random Forest: RF) และเทคนิค (Deep Learning: DL) จากการทดสอบพบว่า เทคนิคที่ดีที่สุดคือเทคนิค (Decision Tree: DT C4.5) มีค่าความถูกต้องมากที่สุด ส่วน Aboalnaser และAlmohammadi [10] ได้ทำการศึกษาแบบจำลองเพื่อการพยากรณ์โรคเบาหวาน ด้วยเทคนิค (Naïve Bayes: NB) เทคนิค (K-Nearest Neighbors: 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 Neighbors: K-NN) มีประสิทธิภาพดีที่สุด ส่วนอัจฉราภรณ์ สุขเพิ่ม และคณะ [11] ได้ทำการสร้างแบบจำลองการวินิจฉัยอัตโนมัติสำหรับความเสี่ยงต่อการเกิดลิ่มเลือกอุดตันในหลอดเลือดดำตามอาการ โดยอาศัยการเรียนรู้ของเครื่อง จากฐานข้อมูลโรงพยาบาลจุฬาลงกรณ์สภากาชาติไทย ด้วยเทคนิค (Decision Tree: DT) เทคนิค (Logistic Regression: LR) และเทคนิค (Neural Network: NN) พบว่าเทคนิคที่ดีที่สุดคือ เทคนิค (Decision Tree: DT) ซึ่งมีค่าความถูกต้องสูงที่สุด และบรรจบ ดลกุล และคณะ [12] ได้ทำการสร้างแบบจำลองเพื่อการพยากรณ์การเกิดแผลที่เท้าของผู้ป่วยโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูล จากฐานข้อมูลหน่วยงานภาครัฐแห่งหนึ่งโดยใช้เทคนิค (Artificial Neural Network: ANN) เทคนิค (Decision Tree: DT) เทคนิค (Support Vector Machine: SVM) พบว่าเทคนิคที่ดีที่สุดคือเทคนิค (Decision Tree: DT) จะเห็นได้ว่าแต่ละเทคนิคที่ผู้วิจัยนำมาศึกษาทดลองแต่ละเทคนิคจะมีประสิทธิภาพที่ไม่แน่นอน ในข้อมูลที่แตกต่างกัน

การคัดเลือกปัจจัย คือขั้นที่ตอนสำคัญขั้นตอนหนึ่งในการทำเหมืองข้อมูลเพื่อลดปัจจัยที่ไม่สำคัญต่อการพยากรณ์ มีนักวิจัยหลายท่านได้นำเอาการคัดเลือกปัจจัยวิธีต่างๆ มาคัดเลือกเพื่อเพิ่มประสิทธิภาพของแบบจำลอง ตัวอย่างเช่นงานวิจัยของ Sujan Ray และคณะ [13] ได้ทำการคัดเลือกคุณสมบัติด้วย Chi-Squared สำหรับโรคหลอดเลือดสมอง วัดประสิทธิภาพด้วยเทคนิค Decision Tree พบว่าค่าความแม่นยำเพิ่มขึ้นจากเดิม 89% เป็น 96.80% คิดเป็นร้อยละ 8.80% รวมถึง สุวิมล สิทธิชาติ [14] ได้ทำการวิเคราะห์คุณลักษณะพื้นฐานทางการศึกษาด้วยเทคนิคเหมืองข้อมูล ทำการคัดเลือกคุณลักษณะด้วย Chi-Square และ Gain Ratio วัดประสิทธิภาพด้วยเทคนิค Multi-layer perceptron และเทคนิค Decision Tree C4.5 พบว่าค่าความถูกต้องเพิ่มขึ้นจากเดิมหลังจากคัดเลือกคุณลักษณะด้วย Chi-Square ร่วมกับเทคนิค Decision Tree c4.5 จากเดิม 66.23% เป็น 80.13% คิดเป็นร้อยละ 13.90%

แบบจำลองที่ได้สร้างขึ้นนี้สามารถช่วยสนับสนุนและเป็นฐานในการตัดสินใจในการรักษาของแพทย์ หรือผู้ที่รับผิดชอบในการดูแลผู้ป่วยโรคเบาหวานและโรคความดันโลหิตสูง ได้ทราบถึงประสิทธิภาพของการรักษาโดยการให้ยาในการรักษาผู้ป่วยโรคเบาหวานและโรคความดันโลหิตสูงในเทคนิคที่ดีที่สุด และยังสามารถนำเทคนิคที่ดีที่สุดไปพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูงในระดับโรงพยาบาลขนาดต่างๆที่มีคนไข้ในกลุ่มนี้ได้ ยิ่งไปกว่านั้นยังสามารถนำข้อมูลที่

พยากรณ์การรักษาของโรคเบาหวานและโรคความดันโลหิตสูงมาวิเคราะห์หาแนวทางการป้องกันหรือการรักษาของแพทย์ได้

ดังนั้นงานวิจัยนี้ผู้วิจัยมีความสนใจที่จะศึกษาประสิทธิภาพของเทคนิค (Decision tree: DT C4.5), (K-Nearest Neighbors: K-NN), (Random Tree: RT), (Multi-layer perceptron: MLP) และ (Logistic Model Trees: LMT) คัดเลือกปัจจัยด้วยวิธี Chi-Square Test, Gain Ratio และ Information Gain ในการสร้างแบบจำลองในการพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูง ของข้อมูลผู้ป่วยโรคเบาหวานและโรคความดันโลหิตสูง จากฐานข้อมูลโรงพยาบาลทุ่งเขาหลวง อำเภอทุ่งเขาหลวง จังหวัดร้อยเอ็ด และทดสอบค่าความไว (Sensitivity) ค่าความจำเพาะ (Specificity) และค่าความถูกต้อง (Accuracy)

1.2 วัตถุประสงค์

1.2.1 เพื่อศึกษาการคัดเลือกปัจจัยในการสร้างแบบจำลองการพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูง

1.2.2 เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูง

1.3 ขอบเขตการวิจัย

1.3.1 รวบรวมข้อมูลจากข้อมูลการรักษาโรคเบาหวานและโรคความดันโลหิตสูง จากฐานข้อมูลโรงพยาบาลทุ่งเขาหลวง อำเภอทุ่งเขาหลวง จังหวัดร้อยเอ็ด ตั้งแต่วันที่ 1 มกราคม 2561 ถึงวันที่ 31 ธันวาคม 2563 (3ปี)

1.3.2 คัดเลือกตัวแปรในการสร้างแบบจำลองการพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูง

1.3.3 สร้างแบบจำลองที่มีประสิทธิภาพในการพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูง

1.3.4 เปรียบเทียบประสิทธิภาพแบบจำลองด้วยวิธีการ 10-fold cross validation

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1.4.1 ได้ความรู้ในการสร้างแบบจำลองพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูง ด้วยเทคนิคต้นไม้ตัดสินใจ เทคนิคเคเนียเรสเนเบอร์ เทคนิคต้นไม้สุ่ม เทคนิคเพอร์เซ็ปตรอนหลายชั้น และเทคนิคลอจิสติกถดถอยแบบต้นไม้

1.4.2 ได้แบบจำลองที่มีประสิทธิภาพในการพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูง เพื่อช่วยสนับสนุนและเป็นฐานในการตัดสินใจในการรักษาของแพทย์ พยาบาล และผู้ที่รับผิดชอบในการดูแลรักษาผู้ป่วยโรคเบาหวานและโรคความดันโลหิตสูงได้เหมาะสมยิ่งขึ้น อีกทั้งยังเป็นแนวทางสำหรับผู้ที่สนใจในการศึกษาเกี่ยวกับการพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูงสำหรับกระทรวงสาธารณสุขในระดับโรงพยาบาลชุมชน โรงพยาบาลทั่วไป หรือโรงพยาบาลที่มีคนไข้โรคเบาหวานและโรคความดันโลหิตสูงต่อไป

1.5 นิยามศัพท์เฉพาะ

1.5.1 โรคเบาหวาน

โรคเบาหวาน เป็นโรคที่เกิดจากความผิดปกติของการทำงานของฮอร์โมนที่ชื่อว่า อินสุลิน (Insulin) ซึ่งโดยปกติแล้วร่างกายของคนเราจำเป็นต้องมีอินสุลิน เพื่อนำน้ำตาลในกระแสเลือดไปเลี้ยงอวัยวะต่างๆของร่างกาย โดยเฉพาะสมองและกล้ามเนื้อ ในภาวะที่อินสุลินมีความผิดปกติ ไม่ว่าจะเป็นการลดลงของปริมาณอินสุลินในร่างกาย หรือการที่อวัยวะต่างๆของร่างกายตอบสนองต่ออินสุลินลดลง (หรือที่เรียกว่า ภาวะดื้ออินสุลิน) จะทำให้ร่างกายไม่สามารถนำน้ำตาลที่อยู่ในกระแสเลือดไปใช้ได้อย่างมีประสิทธิภาพ ทำให้มีปริมาณน้ำตาลคงเหลือในกระแสเลือดมากกว่าปกติ หากน้ำตาลในกระแสเลือดสูงมากขึ้นถึงระดับหนึ่ง จะทำให้ไตซึ่งปกติจะมีหน้าที่ดูดกลับน้ำตาลจากสารที่ถูกกรองจากหน่วยไตไปใช้ ดูดกลับน้ำตาลได้ไม่หมด ส่งผลให้มีน้ำตาลรั่วออกมากับปัสสาวะ จึงเป็นที่มาของคำว่า “โรคเบาหวาน” หากเราปล่อยให้เกิดภาวะเช่นนี้ไปนานๆ โดยไม่ได้รับการรักษาอย่างถูกวิธี จะทำให้เกิดภาวะแทรกซ้อนที่ร้ายแรงตามมาในที่สุด

1.5.2 โรคความดันโลหิตสูง

โรคความดันโลหิต หมายถึง แรงดันของกระแสเลือดที่กระทบต่อผนังหลอดเลือดแดง ซึ่งเกิดจากการสูบฉีดของหัวใจ สามารถวัดได้โดยการใช้เครื่องวัดความดัน (Sphygmomanometer) วัดที่แขน และมีค่าที่วัดได้ 2 ค่า คือ

1.5.2.1 ความดันช่วงบน หรือ ความดันซิสโตลี (Systolic blood pressure) หมายถึง แรงดันเลือดในขณะที่หัวใจบีบตัว ซึ่งอาจจะสูงตามอายุ และความดันช่วงบนของคนคนเดียวกัน อาจมีค่าที่ต่างกันออกไป ตามท่าเคลื่อนไหวของร่างกาย การเปลี่ยนแปลงของอารมณ์ และปริมาณของการออกกำลังกาย

1.5.2.2 ความดันช่วงล่าง หรือ ความดันไดแอสโตลี (Diastolic blood pressure) หมายถึง แรงดันเลือดในขณะที่หัวใจคลายตัว ในปัจจุบันได้มีการกำหนดค่าความดันโลหิตปกติ และระดับความรุนแรงของโรคความดันโลหิตสูง

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ทฤษฎีที่เกี่ยวข้องในบทนี้ประกอบด้วย ประวัติโรงพยาบาล โรคเบาหวาน โรคความดันโลหิตสูง การทำเหมืองข้อมูล เทคนิคเหมืองข้อมูลในการวิจัย และการวัดประสิทธิภาพ และงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 ประวัติโรงพยาบาลทุ่งเขาหลวง

โรงพยาบาลทุ่งเขาหลวง ได้เปิดดำเนินการเมื่อ 1 พฤษภาคม 2556 สังกัดสำนักงานสาธารณสุขจังหวัดร้อยเอ็ด สำนักงานปลัดกระทรวงสาธารณสุข ประเภทโรงพยาบาลขนาดเล็ก ระดับ F3 จำนวน 10 เตียง ประชากรในเขตรับผิดชอบประมาณ 23,573 คน ประกอบด้วย 4807 หลังคาเรือน 51 หมู่บ้าน รับผิดชอบ 5 ตำบล เปิดให้บริการผู้ป่วยนอก และผู้ป่วยใน มีหน่วยบริการปฐมภูมิองค์รวม 1 แห่ง และ รพ.สต.เครือข่าย 5 แห่ง อัตราผู้ป่วยโรคเบาหวานและโรคความดันโลหิตสูงต่อพันประชากร 3 ปีย้อนหลัง ในการเกิดโรคในรายใหม่ ในปี 2561 มีอัตราการเกิดโรคเบาหวานร้อยละ 5.09 ส่วนโรคความดันโลหิตสูงอัตราการเกิดโรคร้อยละ 14.57 ในปี 2562 มีอัตราการเกิดโรคเบาหวานร้อยละ 1.76 ส่วนโรคความดันโลหิตสูงอัตราการเกิดโรคร้อยละ 20.55 และในปี 2563 มีอัตราการเกิดโรคเบาหวานร้อยละ 3.5 และส่วนโรคความดันโลหิตสูงอัตราการเกิดโรคร้อยละ 12.96 จะเห็นได้ว่าอัตราการเกิดโรคสำหรับรายใหม่ยังมีเพิ่มขึ้นในทุกๆปี โดยมีอัตราการเกิดโรคเบาหวานมากที่สุดในปี 2561 และมีอัตราในการเกิดโรคความดันโลหิตสูงมากที่สุดในปี 2562

2.1.2. โรคเบาหวาน

โรคเบาหวาน คือ โรคที่เซลล์ร่างกายมีความผิดปกติในขบวนการเปลี่ยนน้ำตาลในเลือดให้เป็นพลังงาน โดยขบวนการนี้เกี่ยวข้องกับอินซูลินซึ่งเป็นฮอร์โมนที่สร้างจากตับอ่อนเพื่อใช้ควบคุมระดับน้ำตาลในเลือด เมื่อน้ำตาลไม่ได้ถูกใช้จึงทำให้ระดับน้ำตาลในเลือดสูงขึ้นกว่าระดับปกติ โรคเบาหวานแบ่งเป็น 4 ชนิด ตามสาเหตุของการเกิดโรคได้แก่

2.1.2.1 โรคเบาหวานชนิดที่ 1 (Type 1 diabetes mellitus, T1DM) เกิดจากเซลล์ตับอ่อนถูกทำลายจากภูมิคุ้มกันของร่างกาย ทำให้ขาดอินซูลิน มักพบในเด็ก

2.1.2.2 โรคเบาหวานชนิดที่ 2 (Type 2 diabetes mellitus, T1DM) เป็นชนิดที่พบบ่อยที่สุด ร้อยละ 95 ของผู้ป่วยเบาหวานทั้งหมด เกิดจากภาวะดื้อต่ออินซูลิน มักพบในผู้ใหญ่ที่มีน้ำหนักเกินหรืออ้วนร่วมด้วย

2.1.2.3 โรคเบาหวานขณะตั้งครรภ์ (Gestational diabetes mellitus, GDM) เป็นโรคเบาหวานที่เกิดขึ้นขณะตั้งครรภ์ มักเกิดเมื่อไตรมาส 2-3 ของการตั้งครรภ์

2.1.2.4 โรคเบาหวานที่มีสาเหตุจำเพาะ (Specific types of diabetes due to other causes) มีได้หลายสาเหตุ เช่น โรคทางพันธุกรรม โรคของตับอ่อน โรคทางต่อมไร้ท่อ ยาบางชนิด เป็นต้น โดยการวินิจฉัยเบาหวาน ทำได้โดยวิธีใดวิธีหนึ่งใน 4 วิธี ดังต่อไปนี้

2.1.2.4.1 มีอาการโรคเบาหวานชัดเจน ได้แก่ หิวน้ำบ่อย ปัสสาวะบ่อยและปริมาณมาก น้ำหนักตัวลดลงโดยไม่มีสาเหตุ ร่วมกับตรวจระดับน้ำตาลในเลือดเวลาใดก็ได้ ไม่จำเป็นต้องอดอาหาร ถ้ามีค่า ≥ 200 มก./ดล.

2.1.2.4.2 ระดับน้ำตาลในเลือดหลังอดอาหาร (อย่างน้อย 8 ชั่วโมง) ≥ 126 มก./ดล.

2.1.2.4.3 การตรวจความทนต่อกลูโคส โดยให้รับประทานกลูโคส 75 กรัม แล้วตรวจระดับน้ำตาลในเลือดที่ 2 ชั่วโมง ถ้ามีค่า ≥ 200 มก./ดล.

2.1.2.4.4 การตรวจระดับน้ำตาลสะสม (A1C) ≥ 6.5% โดยวิธีการตรวจและห้องปฏิบัติการต้องได้รับการรับรองตามมาตรฐานที่กำหนด ซึ่งยังมีน้อยในประเทศไทย ดังนั้นจึงไม่แนะนำให้ใช้วิธีนี้

ตามแนวทางเวชปฏิบัติของสมาคมโรคเบาหวานแห่งประเทศไทย พ.ศ. 2560 กล่าวว่า การวินิจฉัยโรคเบาหวานตามข้อที่ 2.1.2.4.2-2.1.2.4.4 ต้องมีการตรวจยืนยันอีกครั้งโดยใช้ตัวอย่างเลือดอันใหม่ ด้วยวิธีเดียวกันหรือต่างกันในวันถัดไป อย่างไรก็ตามแนวทางเวชปฏิบัติของสมาคมโรคเบาหวานแห่งสหรัฐอเมริกา พ.ศ. 2562 ได้มีการเปลี่ยนแปลงว่า สามารถตรวจยืนยันอีกครั้งโดยใช้ตัวอย่างเลือดอันเดิมหรืออันใหม่ก็ได้ เพื่อให้การวินิจฉัยโรคเบาหวาน

การรักษาโรคเบาหวาน เป็นการรักษาที่ต้องอาศัยความร่วมมือทั้งจากแพทย์ พยาบาล โภชนากร และตัวผู้ป่วย ผู้ป่วยต้องตระหนักถึงความสำคัญของการรักษา โดยต้องเข้าใจก่อนว่าโรคเบาหวานเป็นโรคเรื้อรังที่ไม่สามารถรักษาให้หายขาดได้ แต่สามารถควบคุมให้อยู่ในเกณฑ์ที่ใกล้เคียงปกติที่สุดได้ ผู้ป่วยเบาหวานสามารถใช้ชีวิตประจำวันและทำกิจกรรมต่างๆ รวมถึงทำงานประจำได้ตามปกติหากแต่ต้องควบคุมระดับน้ำตาลให้อยู่ในเกณฑ์ปกติโดยการควบคุมอาหาร เช่น

ผลไม้ แนะนำให้รับประทานผลไม้ชนิดหวานน้อย การออกกำลังกาย เช่น วิ่งเหยาะๆ ว่ายน้ำ ปั่นจักรยาน และใช้ยาลดระดับน้ำตาลในเลือด แพทย์จะพิจารณาจากชนิดของโรคเบาหวาน เช่น เบาหวานชนิดที่ 1 ควรรักษาโดยการฉีดอินสุลินเท่านั้น ส่วนในเบาหวานชนิดที่ 2 แพทย์จะพิจารณาตามความรุนแรงของโรคและภาวะแทรกซ้อน รวมทั้งการติดตามการรักษาอย่างสม่ำเสมอก็จะช่วยลดโอกาสการเกิดภาวะแทรกซ้อนต่างๆ ที่กล่าวมาแล้วข้างต้นได้

2.1.3 โรคความดันโลหิตสูง

โรคความดันโลหิตสูง คือภาวะที่มีการเปลี่ยนแปลงความดันในหลอดเลือดที่สูงขึ้น ซึ่งในปัจจุบันแพทย์วินิจฉัยว่าในภาวะปกติผู้ที่มีความดันเท่ากับหรือมากกว่า 140/90 มิลลิเมตรปรอท เป็นผู้ที่มีความดันโลหิตสูง หากปล่อยทิ้งไว้ให้อยู่ในระดับนี้นานๆ อาจทำให้อวัยวะต่างๆในร่างกายเสื่อม เช่น มีโอกาสเป็นโรคหัวใจตีบตัน 3-4 เท่า และโรคหลอดเลือดสมองตีบตัน 7 เท่าของผู้ที่มีความดันปกติ และถ้าปล่อยทิ้งไว้ความดันจะเพิ่มขึ้นทุกๆ 10 มิลลิเมตรปรอทต่อปี โรคความดันโลหิตสูงสามารถแบ่งออกเป็น 2 ชนิดได้แก่

2.1.3.1 ชนิดที่ไม่ทราบสาเหตุชัดเจน ส่วนใหญ่เชื่อกันว่าเกิดจาก 2 ปัจจัย คือ กรรมพันธุ์หรือสิ่งแวดล้อม โดยส่วนใหญ่เกิดจากปัจจัยส่งเสริม เช่น ภาวะอ้วน เบาหวาน ทานอาหารเค็ม ดื่มสุรา สูบบุหรี่

2.1.3.2 ชนิดที่ทราบสาเหตุ เป็นโรคความดันโลหิตสูง ซึ่งเป็นผลมาจาก ที่เป็นโรคอื่นมาก่อนและมักต้องรักษาโรคที่เป็นสาเหตุด้วย เช่น เนื้องอกที่ต่อมหมวกไต ไตวายเรื้อรัง หลอดเลือดแดงที่ไปเลี้ยงไตตีบ หลอดเลือดแดงใหญ่ตีบในส่วนของช่องอกรวมถึงผู้ที่ใช้ยาสเตียรอยด์เป็นประจำ ซึ่งมักพบในผู้ป่วยที่มีอายุน้อยกว่า 45 ปี

การวินิจฉัยโรคความดันโลหิตสูง วินิจฉัยได้จากการวัดความดันโลหิตเท่านั้น โดยค่าความดันโลหิตที่วัดที่โรงพยาบาล หากมากกว่า 140 / 90 มิลลิเมตรปรอท จะจัดว่ามีความดันโลหิตสูง การรักษาโรคความดันโลหิตสูงที่เป็นการรักษามาตรฐานมี 2 วิธี คือการปรับเปลี่ยนวิถีการดำเนินชีวิตและการให้ยาลดความดันโลหิต การปรับเปลี่ยนพฤติกรรมชีวิตในระยะยาวเป็นหัวใจสำคัญของการป้องกันกลุ่มโรคไม่ติดต่อ (Non-communicable diseases, NCDs) รวมทั้งโรคความดันโลหิตสูง และยังเป็นพื้นฐานการ ควบคุมความดันโลหิตสำหรับผู้ป่วยโรคความดันโลหิตสูง เช่น การลดน้ำหนักในผู้ที่มีน้ำหนักเกินหรืออ้วน การปรับรูปแบบของการบริโภคอาหารและจำกัดปริมาณเกลือและโซเดียมในอาหาร งดเครื่องดื่มแอลกอฮอล์ การเริ่มให้ยาลดความดันโลหิตในผู้มีโรคความดันโลหิตสูงจะพิจารณาจากปัจจัยหลัก 4 ประการได้แก่ ระดับความดันโลหิตเฉลี่ยที่วัดได้จากสถานพยาบาล ระดับความเสี่ยงที่จะเกิดโรคทางระบบหัวใจและหลอดเลือดของแต่ละบุคคล มีโรคร่วมโดยเฉพาะโรคทางระบบหัวใจและหลอดเลือด และปัจจัยสุดท้ายคือ สภาพของการเกิด target organ damage (TOD)

2.1.4 การทำเหมืองข้อมูล

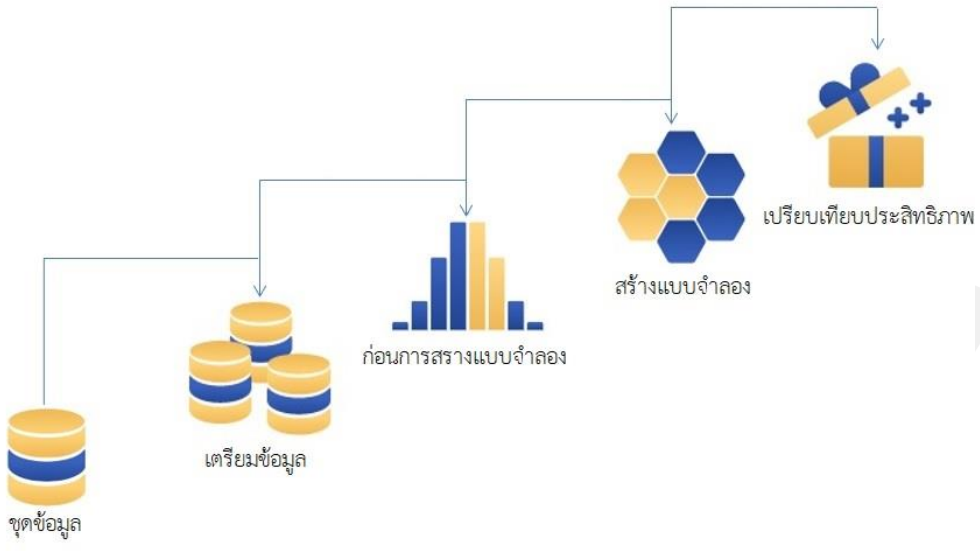
การทำเหมืองข้อมูล (Data Mining) คือกระบวนการค้นหาความรู้ที่อยู่ในฐานข้อมูลขนาดใหญ่ (Knowledge Discovery from very large Databases หรือ KDD) โดยนำข้อมูลที่มีอยู่มาวิเคราะห์แล้วดึงความรู้ หรือสิ่งที่สำคัญออกมาเพื่อใช้ในการวิเคราะห์ หรือพยากรณ์สิ่งต่างๆที่จะเกิดขึ้น ในการค้นหาความรู้นั้นจะหาค่าความจริงที่อยู่ในชุดข้อมูล (Knowledge Discovery) มีขั้นตอนการทำงานดังนี้ [8]

2.1.4.1 การเตรียมข้อมูล คือขั้นตอนการคัดเลือกเฉพาะข้อมูลที่มีอยู่ในฐานข้อมูลที่ต้องการ แล้วนำมาวิเคราะห์ ทำการลบข้อมูลที่ซ้ำซ้อน และข้อมูลที่เป็นค่าว่างออก

2.1.4.2 ขั้นตอนก่อนการสร้างแบบจำลอง คือขั้นตอนการทำความสะอาดข้อมูลที่ไม่สมบูรณ์ รวมถึงการเลือกตัวแปรที่เป็นอิสระจากตัวแปรตาม แล้วแปลงรูปแบบข้อมูลให้อยู่ในรูปแบบที่พร้อมจะนำไปวิเคราะห์

2.1.4.3 การสร้างแบบจำลอง คือกระบวนการในการนำเทคนิคต่างๆในเหมืองข้อมูลมาทำการวิเคราะห์ข้อมูล เพื่อหารูปแบบในการจำแนก เพื่อใช้ในการพยากรณ์

2.1.4.4 การวัดประสิทธิภาพของแบบจำลอง คือการทดลองเพื่อหาค่าความไว (Sensitivity) ค่าความจำเพาะ (specificity) และค่าความถูกต้อง (Accuracy)



ภาพที่ 2.1 ขั้นตอนในการทำเหมืองข้อมูล

2.1.5 เทคนิคเหมืองข้อมูลในการวิจัย

2.1.5.1 เทคนิคต้นไม้ตัดสินใจ

เทคนิคต้นไม้ตัดสินใจ แบบ C4.5 (Decision tree: DT C4.5) เป็นวิธีการในการสร้างแบบจำลองโดยใช้สูตรทางคณิตศาสตร์ ได้พัฒนามาจากวิธีพื้นฐานคือวิธี ID3 เพื่อทำการคัดเลือกหาปัจจัยที่ดีที่สุด [15] เรียนรู้โดยการจำแนก (Classification) ชุดข้อมูลออกเป็นกลุ่ม (Class) ต่างๆ เพื่อการตัดสินใจว่าข้อมูลในชุดนี้ควรจัดอยู่ในกลุ่มใด โดยพิจารณาจากค่าความแตกต่างใน (Entropy) ซึ่งโมเดลที่ใช้จะเลือกคุณลักษณะออกเป็นกลุ่มๆตามที่กำหนดไว้ด้วยค่า (Information Gain) ซึ่ง C4.5 จะทำงานแบบ (Supervised Learning) สามารถสร้างแบบจำลองจากตัวอย่างของข้อมูลได้ก่อนล่วงหน้า (Training Set) และพยากรณ์กลุ่มข้อมูลที่ยังไม่ได้จัดหมวดหมู่ได้ด้วย ต้นไม้ตัดสินใจจะใช้วิธีการหาตัวแปร (Attribute) ที่สำคัญที่สุดนำมาแบ่งข้อมูลแล้วนำมาสร้างเป็นโหนดราก (Root Note) โดยมีตัวแปรตาม (Target Attribute) เป็นโหนดสุดท้าย (Leaf Node) ที่ถูกกำหนดไว้ก่อนแล้วนำค่าที่เป็นไปได้ในตัวแปรต้น (Independent Attribute) ที่ถูกเลือกมาแบ่งกลุ่มข้อมูลที่แตกออกจากโหนดราก (Root Node) และหาค่าตัวแปรต้น (Independent Attribute) ที่สำคัญที่สุดจากข้อมูลที่เข้ามาเพื่อหาตัวแบ่งต่อไป [16] โดยคำนวณหาค่า Gain ratio จาก Information Gain และ Spit info [17] โดยการวัด Gain Ratio ในแต่ละคลาสการคำนวณ Gain ratio โดยใช้ค่า Spit info ในสมการที่ 2.1 และการคำนวณค่าวัด Gain Ratio ดังสมการที่ 2.2 [18]

$$Spitinfo = \sum_{i=1}^{k} \left(\frac{ni}{n} \log 2 \frac{ni}{n} \right) \tag{2.1}$$

$$GainRatio = \frac{Info}{SplitInfo} \tag{2.2}$$

คำนวณค่า Gain สำหรับแต่ละมิติข้อมูล มิติข้อมูลใดมีค่า Gain สูงสุด จะถูกเลือกให้เป็นกลุ่มย่อยที่สามารถจำแนก ดังสมการที่ 2.3 แสดงการคำนวณค่า Entropy และคำนวณค่า Gain ดังสมการที่ 2.4

$$Entropy(p) = -\sum_{i=0}^{c-1} p(j \mid t) \log_2 p(j \mid t) \tag{2.3}$$

โดยที่ $\sum_i$ คือผลรวมของความน่าจะเป็นของค่า j ทีเกิดในคลาส t

$p(j \mid t)$ คือค่าความถีทีมีความสัมพันธ์ของกลุ่ม j กับโหนด t

$$Gain = Entropy(p) - \left(\sum_{i=0}^{k} \frac{ni}{n} Entropy(i) \right) \quad (2.4)$$

โดยที่ $Entropy(p)$ คือค่า Entropy ของตัว Root

$\left(\sum_{i=0}^{k} \frac{ni}{n} Entropy(i) \right)$ คือ ค่า Entropy ในแต่ละโหนดย่อย

2.1.5.2 เทคนิคเคเนียเรสเนเบอร์

เทคนิคเคเนียเรสเนเบอร์ (K-Nearest Neighbor: KNN) เป็นเทคนิคในการสร้างแบบจำลองเพื่อจำแนกชุดข้อมูล (Data Classification) จะจำแนกประเภทข้อมูลแบบ (Supervised Learning) การเรียนรู้แบบมีการสอน จากนั้นใช้โมเดลจำแนกประเภทของข้อมูลจะใช้วิธีการวิเคราะห์จากข้อมูลที่มีคุณสมบัติใกล้เคียงที่สุดจำนวน k ตัว กับข้อมูลที่ต้องการจำแนกประเภทของข้อมูลที่ต้องการทำนาย กลุ่มของข้อมูลชุดใหม่ โดยจะทำการทำนายตามคลาสส่วนใหญ่ของข้อมูล [19] การกำหนดค่า k ให้กับ KNN จะเป็นการกำหนดขอบเขตในการวิเคราะห์ข้อมูล คำนวณระยะห่าง (Distance) ของข้อมูลที่ต้องการจำแนกตามจำนวน k ที่กำหนดไว้ และจัดเรียงลำดับของระยะห่าง แล้วพิจารณาชุดข้อมูลที่ใกล้จุดที่ต้องการตามจำนวน k และสังเกตว่า (Class) ไหนใกล้จุดที่พิจารณาเป็นจำนวนมากที่สุด ซึ่งถ้ากำหนดค่า k ไว้สูงก็จะส่งผลให้ใช้เวลาในการประมวลผลนานขึ้น KNN ทำให้เราทราบถึงประเภทข้อมูลของสิ่งที่เรายังไม่เคยจำแนกมาก่อนได้ [20] การหาความยาวระหว่างจุดที่ต้องการโดยใช้เครื่องมือและวิธีต่าง ๆ งานวิจัยได้เลือกวิธีการหามาตรวัดความแม่นยำ โดย Euclidean Distance ระยะทาง ระหว่าง 2 จุดเพราะง่ายต่อความเข้าใจ และลักษณะการคำนวณที่คล้ายกับทฤษฎีบทพีทาโกรัส ดังภาพที่ 2.2 โดยกำหนดให้มีจุด p อยู่ที่พิกัดในแนวแกน x และ y เป็น p= (p1,p2,p3,...,pn) และจุดอยู่ที่พิกัด q= (q1,q2,q3,...,qn) การคำนวณระยะห่าง ระหว่างจุด p และ q ด้วยวิธีการ Euclidean Distance สามารถแสดงขั้นตอนการคำนวณได้ดังสมการที่ 2.5 [21]

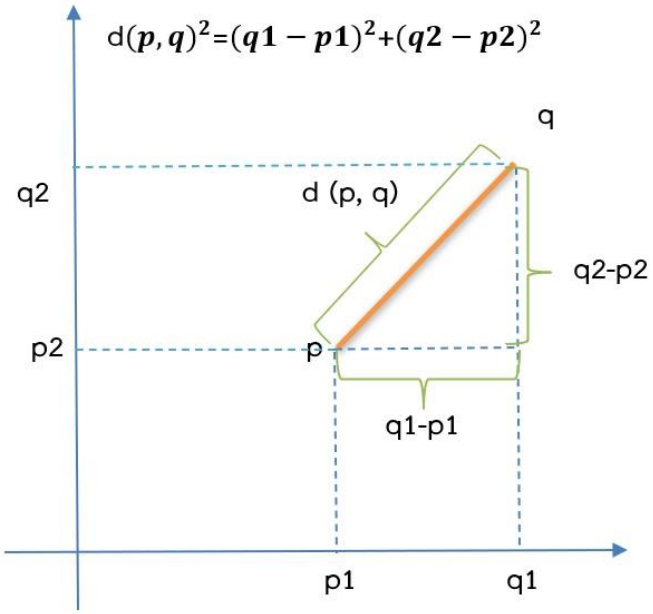
$$d1(q,p) = \sqrt{\sum_{i=1}^{n} \left(qi - pi \right)^2} \quad (2.5)$$

โดยที่ $d_1(q, p)$ คือระยะทางจากจุด p ไปยังจุด q วัดในแบบ Euclidean

p คือจุดใดๆ

q คือจุดใดๆ

n คือจำนวนมิติของข้อมูล



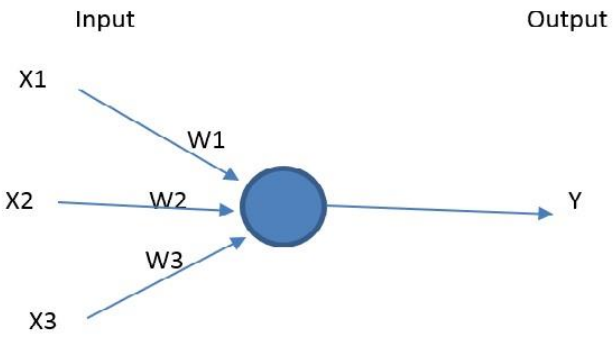
ภาพที่ 2.2 ตัวอย่างที่ใช้ในการคำนวณหาระยะทาง

2.1.5.3 เทคนิคต้นไม้สุ่ม

เทคนิคต้นไม้สุ่ม (Random Tree: RT) คือเทคนิคที่ใช้ในการจำแนกหมวดหมู่ เช่นเดียวกับ C4.5 โดยมีหลักการสร้าง Tree จากการสุ่ม Tree หลายๆแบบ ในแต่ละโหนดแล้วเลือกมาประมวลผลโดยไม่ใช้การ Prune และเนื่องจากจำนวนของ Tree เพิ่มขึ้นอย่างรวดเร็ว ซึ่งยากแก่การแก้ปัญหา การสร้าง Tree ที่เป็นไปได้ทั้งหมด ส่วนประกอบสามารถสร้างชุดสุ่มของ Tree ออกมาจากการกระจายชุดต่างๆของ Tree Random Tree เป็นการสุ่มวาดที่สุ่มจากชุดของ Tree ที่เป็นไปได้ ในบริบทนี้ “สุ่ม” หมายความว่า Tree ในชุดของ Tree แต่ละ Tree มีโอกาสเท่าเทียมกันของการเป็นตัวอย่าง วิธีที่บอกนี้ก็คือ ว่าการกระจายของ Tree คือ “ชุด” Tree สุ่มแบบต่อเนื่อง (CRT) เป็นแบบสุ่ม Tree จริง Tree และมีระยะในการเดินทาง สั้นที่สุด Tree ชุดย่อย ประกอบด้วย Tree ที่มี n จุด แต่ละองค์ประกอบของโครงสร้างที่แท้จริงคือราก [22]

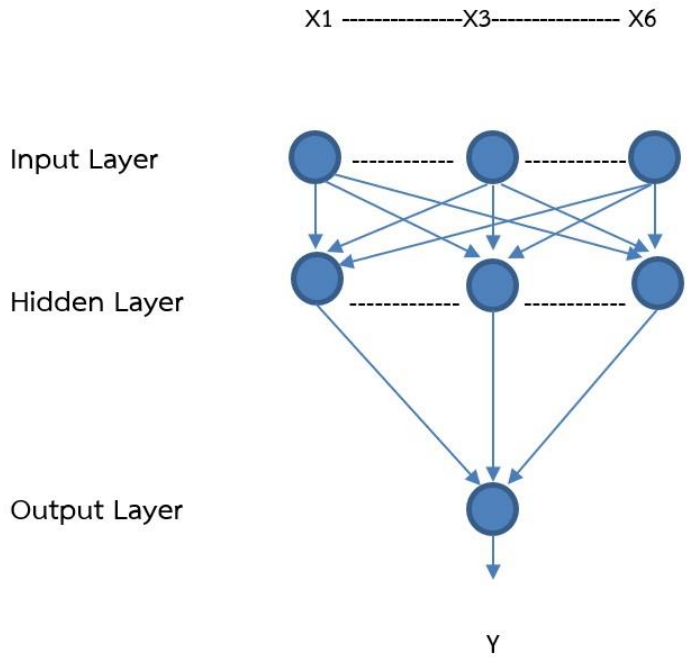
2.1.5.4 เทคนิคเพอร์เซ็ปตรอนหลายชั้น

เทคนิคเพอร์เซ็ปตรอนหลายชั้น (Multi-layer perceptron: MLP) คือเทคนิคในกลุ่มเทคนิคโครงข่ายประสาทเทียม เทคนิคนี้ใช้กระบวนการทางคณิตศาสตร์ที่จำลองมาจากการทำงานของระบบสมองของมนุษย์เพื่อนำมาใช้ประโยชน์ในการตัดสินใจ การทำนาย การจำแนกรูปแบบของโครงข่ายประสาทเทียม พื้นฐานจะประกอบด้วยเซลล์ประสาท (Neuron) เพียง 1 เซลล์ ซึ่งเราเรียกว่าสถาปัตยกรรมเพอร์เซปตรอน (Perceptron Architecture) โดยที่อินพุตต่างๆ จะถูกส่งผ่านกลุ่มใยประสาทนำเข้า (Dendrites) เข้าสู่เซลล์ประสาทเพื่อทำการประมวลผลและส่งผลลัพธ์ออกมาทางแกนประสาท (Axon) โดยการประมวลผลจะใช้ผลรวมของผลคูณ (Sum of Product) ของข้อมูลอินพุตและค่าน้ำหนัก (Weight) ที่ใยประสาทนำเข้า ตัวอย่างของรูปแบบสถาปัตยกรรมเพอร์เซปตรอน แสดงดัง ภาพที่ 2.3 โดยข้อมูล x1 ถึง x3 คือข้อมูลอินพุต w1 ถึง w3 คือ ค่าน้ำหนัก และ y คือเอาต์พุตที่ได้จากการประมวลผลของเซลล์ประสาท [23]



ภาพที่ 2.3 โครงข่ายเพอร์เซปตรอน

ในงานวิจัยนี้ได้ใช้สถาปัตยกรรมเพอร์เซปตรอนแบบหลายชั้น [24] (Multi-Layer Perceptron : MLP) โครงข่ายประสาทเทียมแบบ (MLP) เป็นรูปแบบหนึ่งของโครงข่ายประสาทเทียมที่มีโครงสร้างเป็นแบบหลายๆชั้น ใช้สำหรับงานที่มีความซับซ้อน โดยมีกระบวนการฝึกฝนเป็นแบบมีผู้สอน (Supervised Learning) และใช้ขั้นตอนการส่งค่าย้อนกลับ (Back-propagation) ซึ่งวิธีการนี้สามารถแก้ปัญหาที่มีความซับซ้อน ได้อย่างมีประสิทธิภาพ ประกอบไปด้วยชั้นอินพุต (Input Layer) ชั้นซ่อน (Hidden Layer) และ ชั้นเอาต์พุต (Output Layer) แสดงดังภาพที่ 2.4



ภาพที่ 2.4 โครงข่ายเพอร์เซปตรอนแบบหลายชั้น

ข้อมูลที่เข้ามาจะถูกส่งไปคำนวณในชั้นซ่อนเพื่อหาผลรวมของผลคูณของข้อมูลเข้าและค่าน้ำหนัก แสดงดังสมการที่ 2.6

$$y = \sum_{i=0}^{n} xi, w_i \qquad (2.6)$$

y คือค่าผลรวมของผลคูณข้อมูลเข้า x_i และน้ำหนัก w_i

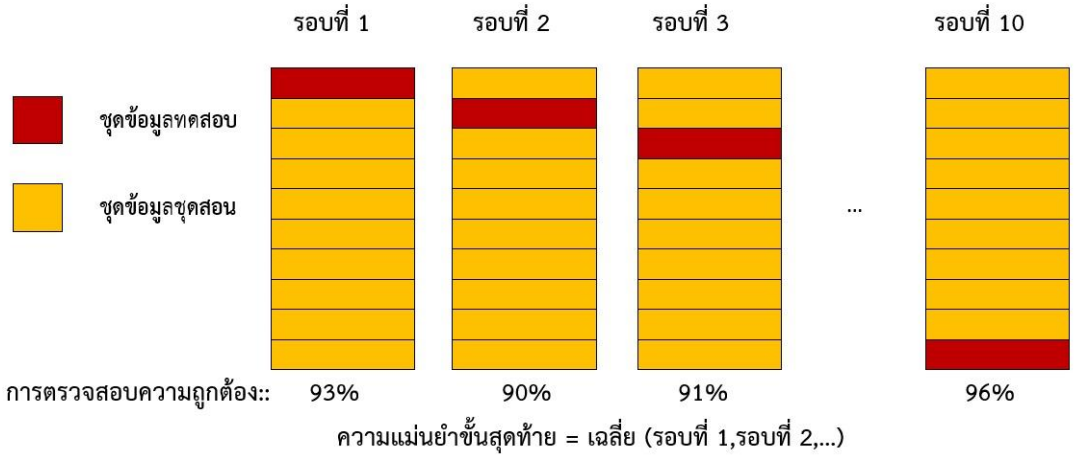
i คือจำนวนข้อมูลเข้าหรือจำนวนค่าน้ำหนัก

2.1.5.5 เทคนิคลอจิสติกถดถอยแบบต้นไม้

เทคนิคลอจิสติกถดถอยแบบต้นไม้ (Logistic Model Trees: LMT) [25] คือการรวมกันของเทคนิคการถดถอยลอจิสติก (Logistic Regression) และเทคนิคต้นไม้ (Trees) โครงสร้างแบบลอจิสติกเป็นโครงสร้างการตัดสินใจใช้ฟังก์ชันถดถอยที่ใบเหมือนต้นไม้จำลอง การทดสอบคุณลักษณะคือเชื่อมโยงทุกโหนดภายในสำหรับแอตทริบิวต์ที่ระบุค่า K และอินสแตนซ์จะถูกจัดเรียงในค่า k และจะขึ้นอยู่กับแอตทริบิวต์ด้วย

2.1.6 การวัดประสิทธิภาพ

การวัดประสิทธิภาพ (Evaluation) เป็นขั้นตอนการประเมินผลลัพธ์ว่ามีความเหมาะสมหรือตรงกับวัตถุประสงค์ที่ต้องการหรือไม่ซึ่งควรนำเสนอผลการวิเคราะห์ในรูปแบบที่ผู้ใช้งานสามารถเข้าใจได้ง่าย วัดประสิทธิภาพของแบบจำลองโดยใช้เทคนิคการวัดประสิทธิภาพแบบ 10-fold cross validation โดยการแบบข้อมูลออกเป็น 10 กลุ่มเท่าๆ กันโดยในแต่ละรอบการทดสอบจะใช้ข้อมูล 1 ชุด เป็นชุดทดสอบและใช้ชุดที่เหลือเป็นชุดฝึกสอน และในการทดลองครั้งที่สองจะใช้ข้อมูลชุดที่ 2 เป็นชุดข้อมูลทดสอบและให้ข้อมูลชุดที่เหลือเป็นข้อมูลชุดฝึกสอน ทำจนกระทั่งข้อมูลทุกชุดข้อมูลได้ถูกนำมาเป็นชุดข้อมูลทดสอบทั้งหมด ซึ่งจำนวนในการทดสอบมีจำนวนเท่ากับ K ครั้ง โดยผลลัพธ์ที่ได้นั้นจะมาคำนวณหาค่าเฉลี่ยความถูกต้องของการจำแนกข้อมูลในแต่ละรอบ



ภาพที่ 2.5 ตัวอย่างการทดสอบประสิทธิภาพแบบ 10- fold cross validation

จากภาพที่ 2.5 จะเป็นการทดสอบประสิทธิภาพแบบ 10-fold ซึ่งจะทำการแบ่งชุดข้อมูลออกเป็น 10 ชุด โดยในแต่ละรอบจะใช้ชุดข้อมูลเพื่อเป็นชุดข้อมูลทดสอบ 1 ชุด และให้ชุดข้อมูลอื่นๆ เป็นข้อมูลชุดสอน โดยจะทำการทดสอบทั้งหมด 10 รอบ ในการวัดประสิทธิภาพการทำงานในแต่ละขั้นตอนวิธี สามารถวัดได้จากผลของการจำแนกกลุ่มของข้อมูลและสามารถหาค่าความไว (Sensitivity) ค่าความจำเพาะ (specificity) และค่าความถูกต้อง (Accuracy) ดังสมการ

2.1.6.1 ค่าความไว (Sensitivity) คือค่าที่แบบจำลองสามารถพยากรณ์ข้อมูลผู้ป่วยที่เกิดโรคได้อย่างถูกต้องต่อผู้ป่วยที่เกิดโรคจริง ดังสมการที่ 2.7

$$Sensitivity = \frac{TN}{TN + FN} \tag{2.7}$$

2.1.6.2 ค่าความจำเพาะ (Specificity) คือค่าที่แบบจำลองสามารถพยากรณ์ข้อมูลผู้ป่วยที่ไม่เกิดโรคได้อย่างถูกต้องต่อผู้ป่วยที่พยากรณ์ว่าเกิดโรค ดังสมการที่ 2.8

$$Specificity = \frac{TN}{TN + FP} \tag{2.8}$$

2.1.6.3 ค่าความถูกต้อง (Accuracy) คือค่าที่แบบจำลองสามารถจำแนกข้อมูลผู้ป่วยที่เกิดโรค และไม่ เกิดโรคได้อย่างถูกต้องต่อข้อมูลทั้งหมด ดังสมการที่ 2.9

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.9}$$

เมื่อ TP คือ จำนวนข้อมูลที่แบบจำลองพยากรณ์การเกิดโรคได้อย่างถูกต้อง

TN คือ จำนวนข้อมูลที่แบบจำลองพยากรณ์การเกิดโรคได้อย่างถูกต้อง

FP คือ จำนวนข้อมูลที่แบบจำลองพยากรณ์การเกิดโรคได้ไม่ถูกต้อง

FN คือ จำนวนข้อมูลที่แบบจำลองพยากรณ์การไม่เกิดโรคได้ไม่ถูกต้อง

2.2 งานวิจัยที่เกี่ยวข้อง

นงเยาว์ ในอรุณ [26] ได้ทำการสร้างแบบจำลองทำนายความเสี่ยงการเกิดโรคหัวใจและหลอดเลือดจากข้อมูลผู้ป่วยที่มารับบริการสาธารณสุขในจังหวัดสระบุรี ด้วยเทคนิคโครงข่ายประสาทเทียม (Neural Networks: NN) เทคนิคฟอเรสต์แบบสุ่ม (Random Forest: RF) เทคนิคเค-เนียร์เรสเนเบอร์ (K-Nearest Neighbors: K-NN) เทคนิคนาอีฟเบย์ (Naïve Bayes: NB) และเทคนิคต้นไม้ตัดสินใจ (Decision Tree: DT) โดยเปรียบเทียบประสิทธิภาพของเทคนิคด้วยค่าความถูกต้อง (Accuracy) และค่าความคลาดเคลื่อน (Error) พบว่าเทคนิคโครงข่ายประสาทเทียม (Neural Networks: NN) สามารถสร้างแบบจำลองได้ค่าความถูกต้องมากที่สุด

เอพร โมพี และคณะ [27] ได้ทำการเสนอความสัมพันธ์ของปัจจัยที่ส่งผลต่อการเกิดโรคเรื้อรัง และรูปแบบสภาวะเสี่ยงโรคเรื้อรังของผู้สูงอายุ อำเภอเมืองสุราษฎร์ธานี จังหวัดสุราษฎร์ธานี ด้วยเทคนิค Decision Tree: C4.5 เทคนิค Decision Rule: Partial Rules เทคนิค Decision Tree: Induction Association Rule เปรียบเทียบประสิทธิภาพด้วยค่าความแม่นยำ (Correctly) ค่าความถูกต้อง (Precision) ค่าระลึก (Recall) และค่าความเหวี่ยง (F-measure) พบว่า เทคนิค Decision Tree: C4.5 มีค่าความถูกต้องมากที่สุด

อุกฤษฎ์ ศรีสุข และจารี ทองคำ [9] ได้ทำการนำเสนอการสร้างแบบจำลองเพื่อการพยากรณ์การเกิดโรค โรคมะเร็งเต้านม โรคเบาหวาน โรคไฮเปอร์ไทรอยด์ จากฐานข้อมูล University of California (UCI) ด้วยเทคนิคต้นไม้ตัดสินใจ (Decision Tree: DT C4.5) เทคนิคนาอีฟเบย์ (Naïve Bayes: NB) เทคนิคโครงข่ายประสาทเทียม (Neural Networks: NN) เทคนิคป่าสุ่ม (Random Forest: RF) และเทคนิคการเรียนรู้อย่างลึกซึ้ง (Deep Learning: DL) เปรียบเทียบประสิทธิภาพด้วยค่าความถูกต้อง (Accuracy) ค่าความไว (Sensitivity) และค่าจำเพาะ (Specificity) จากการทดสอบพบว่า เทคนิคที่ดีที่สุดคือเทคนิคต้นไม้ตัดสินใจ (Decision Tree: DT C4.5) มีค่าความถูกต้องมากที่สุด

Aboalnaser และ Almohammadi [10] ได้ทำการศึกษาการสร้างแบบจำลองพยากรณ์โรคเบาหวาน โดยมีทั้งหมด 4 ขั้นตอนการดำเนินงาน ขั้นตอนที่ 1 รวบรวมฐานข้อมูลจากสถาบันเบาหวาน จำนวน 2,000 ระเบียน ประกอบด้วย 8 ปัจจัย ได้แก่ Pregnancies, Glucose Plasma, Blood, Skin Thickness, BMI, Diabetes, Age, Outcome ขั้นตอนที่ 2 วิเคราะห์ข้อมูลโดยใช้ Remove missing values, Select relevant features, Normalize data ขั้นตอนที่ 3 ทำการสร้างแบบจำลองเพื่อการพยากรณ์ ด้วยเทคนิคนาอีฟเบย์ (Naïve Bayes: NB) เทคนิคเค-เนียร์เรส

เนเบอร์ (K-Nearest Neighbors: K-NN) เทคนิคโครงข่ายประสาทเทียม (Artificial Neural Network: ANN) เทคนิคต้นไม้ตัดสินใจ (Decision Tree: DT) เทคนิคป่าสุ่ม (Random Forest: RF) เทคนิคซับพอร์ตเวกเตอร์แมทชีน (Support Vector Machine: SVM) และเทคนิคการถดถอยโลจิสติก (Logistic Regression: LR) ขั้นตอนที่ 4 วัดประสิทธิภาพโดยการแบ่งข้อมูลเป็นเป็นชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบด้วยหลักการ 10-fold cross-validation จากนั้นวัดค่าด้วยค่าความถูกต้อง(Accuracy) ค่าความถ่วงดุล (F-measure) ค่าความแม่นยำ (Precision) และค่าระลึก (Recall) พบว่า เทคนิคเค-เนียร์เรสเนเบอร์ (K-Nearest Neighbors: K-NN) มีประสิทธิภาพดีที่สุด

Bamal และคณะ [28] เปรียบเทียบประสิทธิภาพของแบบจำลองการพยากรณ์การเกิดโรคเบาหวาน โดยมีทั้งหมด 3 ขั้นตอนการดำเนินงาน ขั้นตอนที่ 1 ทำการรวบรวมข้อมูลจากฐานข้อมูล University of California (UCI) ชุดข้อมูลมี 2,000 ระเบียนและ 9 คุณสมบัติ ระดับผลลัพธ์ที่คาดการณ์ไว้เพื่อที่จะนำมาวิเคราะห์ ได้แก่ การตั้งครรภ์ (pgn) ความดันโลหิตช่วงไดแอสโตลิก (bp mm Hg) ความหนาของผิวหนัง (ผิวหนังหนา มม.) อินซูลิน (In mu U/ml) กลูโคสในพลาสมา (GP) ดัชนีมวลกาย (BMI) ฟังก์ชันสายเลือดเบาหวาน (Dpf) อายุ (ปี) และ พารามิเตอร์คลาสพิเศษ ('ใช่' หรือ 'ไม่ใช่') ขั้นตอนที่ 2 สร้างแบบจำลองด้วยเทคนิคซับพอร์ตเวกเตอร์แมทชีน Support Vector Machine (SVM) เทคนิคนาอีฟเบย์ (Naïve Bayes: NB) เทคนิคต้นไม้ตัดสินใจ (Decision Tree: DT) เทคนิคป่าสุ่ม (Random Forest: RF) เทคนิคถดถอยเชิงเส้น (Linear Regression: LR) เทคนิคเค-เนียร์เรสเนเบอร์ (K-Nearest Neighbors: K-NN) และเทคนิคโครงข่ายประสาทเทียม (Neural Networks: NN) ขั้นตอนที่ 3 ทำการวัดประสิทธิภาพด้วยค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าระลึก (Recall) พบว่าเทคนิคที่มีค่าความถูกต้องมากที่สุดคือเทคนิคป่าสุ่ม (Random Forest: RF)

บทที่ 3
วิธีการดำเนินงานวิจัย

วิธีการดำเนินงานวิจัยในงานวิจัยนี้ได้ใช้ขั้นตอน 4 ขั้นตอน [9] ในการทำเหมืองมูล ประกอบด้วย ขั้นตอนการเตรียมข้อมูล ขั้นตอนก่อนสร้างแบบจำลอง ขั้นตอนการสร้างแบบจำลอง และขั้นตอนการวัดประสิทธิภาพของแบบจำลอง

3.1 การเตรียมข้อมูล

การเตรียมข้อมูลในงานวิจัยนี้ได้นำข้อมูลการรักษาผู้ป่วยโรคเบาหวานและผู้ป่วยโรคความดันโลหิตสูงจากฐานข้อมูลโรงพยาบาลทุ่งเขาหลวง อำเภอทุ่งเขาหลวง จังหวัดร้อยเอ็ด ในช่วงวันที่ 1 มกราคม 2561 ถึง 31 ธันวาคม 2563 ซึ่งเป็นโรคที่ผู้ป่วยมารับการรักษาเป็นอันดับต้นๆของโรงพยาบาล โดยใช้โปรแกรม HosXP ค้นหาข้อมูลที่เกี่ยวข้อง ด้วยชุดคำสั่งโปรแกรมภาษา SQL และชุดคำสั่งภาษา PHP เพื่อประมวลผลข้อมูลในการจัดเตรียมคลาสหรือปัจจัยตาม และส่งออกข้อมูลอยู่ในรูปแบบของ Excel ประกอบด้วยชุดข้อมูลที่แตกต่างกันจำนวน 2 ชุดข้อมูลคือ ข้อมูลชุดที่ 1 คือ ข้อมูลผู้ป่วยโรคเบาหวานมีชนิดตัวแปรเป็น Numeric และ Nominal ทั้งหมด 14 ตัวแปร ข้อมูลชุดที่ 2 คือข้อมูลผู้ป่วยโรคความดันโลหิตสูงตัวแปรเป็น Numeric และ Nominal ทั้งหมด 13 ตัวแปร ดังตารางที่ 3.1 และตางรางที่ 3.2

ตารางที่ 3.1 ตัวแปรผู้ป่วยโรคเบาหวาน

ลำดับ	ชื่อตัวแปร	รายละเอียด	ชนิดตัวแปร
1	Sex	เพศ	Nominal
2	Age	อายุ	Numeric
3	Occ	อาชีพ	Nominal
4	Marry	สถานะภาพ	Nominal
5	Edu	ระดับการศึกษา	Nominal
6	Bps	ค่าความดันช่วงบน	Numeric
7	Bpd	ค่าความดันช่วงล่าง	Numeric
8	Pulse	อัตราชีพจร	Numeric
9	Rr	อัตราการหายใจ	Numeric
10	Bmi	ดัชนีมวลกาย	Numeric
11	Dm1	Metformin	Nominal
12	Dm2	Glipizide	Nominal
13	Dm3	มะระขี้นก	Nominal
14	Class	ผลการรักษา (ระดับน้ำตาลปกติ, ระดับน้ำตาลสูง)	Nominal

จากตารางที่ 3.1 แสดงตัวแปรต้นและตัวแปรตามของผู้ป่วยโรคเบาหวานโดยตัวแปรต้นประกอบด้วยข้อมูลพื้นฐานของผู้ป่วยและตัวแปรยาที่ใช้ในการรักษาผู้ป่วย และตัวแปรตามซึ่งเป็นคลาส ตัวแปรตามนี้เป็นค่าน้ำตาลหลังจากการได้รับยาของการรักษาประกอบด้วย ระดับน้ำตาลปกติ (ค่าน้ำตาล ≤126 มิลลิกรัม/เดซิลิตร) และระดับน้ำตาลสูง (ค่าน้ำตาล >126 มิลลิกรัม/เดซิลิตร)

ตารางที่ 3.2 ตัวแปรผู้ป่วยโรคความดันโลหิตสูง

ลำดับ	ชื่อตัวแปร	รายละเอียด	ชนิดตัวแปร
1	Sex	เพศ	Nominal
2	Age	อายุ	Numeric
3	Occ	อาชีพ	Nominal
4	Marry	สถานะภาพ	Nominal
5	Edu	ระดับการศึกษา	Nominal
6	Bpd	ค่าความดันช่วงล่าง	Numeric
7	Pulse	อัตราชีพจร	Numeric
8	Rr	อัตราการหายใจ	Numeric
9	Bmi	ดัชนีมวลกาย	Numeric
10	Ht1	Amlodipine	Nominal
11	Ht2	Enalapril 5 mg	Nominal
12	Ht3	Enalapril 10 mg	Nominal
13	Class	ผลการรักษา (ระดับความดันปกติ, ระดับความดันสูง)	Nominal

จากตารางที่ 3.2 แสดงตัวแปรต้นและตัวแปรตามของผู้ป่วยโรคความดันโลหิตสูงโดยตัวแปรต้นประกอบด้วยข้อมูลพื้นฐานของผู้ป่วยและตัวแปรยาที่ใช้ในการรักษาผู้ป่วย และตัวแปรตามซึ่งเป็นคลาส ตัวแปรตามนี้เป็นค่าความดันโลหิตหลังจากการได้รับยาของการรักษาประกอบด้วย ระดับความดันปกติปกติ (ค่าความดัน ≤140/90 มิลลิเมตรปรอท) และระดับความดันสูง (ค่าความดัน >140/90 มิลลิเมตรปรอท)

3.2 ขั้นตอนก่อนสร้างแบบจำลอง

ขั้นตอนก่อนการสร้างแบบจำลองในงานวิจัยนี้ เริ่มจากการรวบรวมข้อมูลที่เกี่ยวข้อง ทำความสะอาดข้อมูลที่ไม่สมบูรณ์โดยการตัดค่าที่ว่างหรือแทนที่ค่าที่เป็นไปได้ในชุดข้อมูล แล้วจึงทำการแปลงรูปแบบข้อมูลให้อยู่ในรูปแบบที่พร้อมจะนำไปวิเคราะห์ รวมถึงการเลือกตัวแปรอิสระที่มีผลต่อการพยากรณ์การรักษาโรคเบาหวาน และโรคความดันโลหิตสูง 3 วิธี ได้แก่วิธี Chi-Square test, Gain Ratio และ Information Gain

วิธี Chi square test [29] ใช้การทดสอบความสัมพันธ์ระหว่างตัวแปร (Test of association) ดังสมการที่ 3.1

$$Chi^2 = \sum \frac{\left(O-E\right)^2}{E} \qquad (3.1)$$

เมื่อ Chi^2 แทนตัวสถิติทดสอบไคสแควร์

O แทนความถี่ที่ได้จาการสังเกต

E แทนความถี่ที่คาดหวัง

วิธี Gain Ratio [30] เทคนิคการวัดจำนวนบิตของข้อมูล เพื่อใช้ในการทำนายคาดเดาจำแนกประเภทของข้อมูล ดังสมการที่ 3.2

$$Gain = Entropy(p) - \left(\sum_{i=0}^{k} \frac{n_i}{n} Entropy(i) \right) \qquad (3.2)$$

โดยที่

$Entropy(p)$ คือค่า Entropy ของตัว Root

$\left(\sum_{i=0}^{k} \frac{n_i}{n} Entropy(p) \right)$ คือค่า Entropy ในแต่ละโหนดย่อย

วิธี Information Gain [31] คือการพิจารณาจากค่าความน่าจะเป็นของแต่ละคุณลักษณะที่เป็นไปได้แล้ววัดค่าความไร้ระเบียบ (Entropy) เพื่อคัดเลือกคุณลักษณะที่มีความสำคัญในการจำแนกกลุ่ม ดังสมการที่ 3.3

$$Entropy(p) = -\sum_{i=0}^{c-1} p(j \mid t) \log_2 p(j \mid t) \quad (3.3)$$

โดยที่

$\sum i$ คือผลรวมของความน่าจะเป็นของค่า j ทีเกิดในคลาส t

$p(j \mid t)$ คือค่าความถี่ที่มีความสัมพันธ์ของกลุ่ม j กับโหนด t

3.3 ขั้นตอนการสร้างแบบจำลอง

การสร้างแบบจำลอง ในงานวิจัยนี้ได้ใช้โปรแกรม WEKA ซึ่งเป็นโปรแกรมที่มีความนิยมและมีความน่าเชื่อถือในการทดลอง [32] โดยเทคนิคที่นำมาสร้างแบบจำลองเป็นเทคนิคที่มีประสิทธิภาพในการสร้างแบบจำลองการพยากรณ์

1. เทคนิคต้นไม้ตัดสินใจ C4.5 เป็นเทคนิคที่ใช้น้ำหนักของแต่ละปัจจัยมาเป็นโหนดเพื่อสร้างต้นไม้ตัดสินใจ โดยพารามิเตอร์ของ C4.5 คือความเชื่อมั่นของแต่ละปัจจัยที่นำมาเป็นโหนด (Confidence Factor) งานวิจัยนี้ได้กำหนดความเชื่อมั่นของแต่ละปัจจัยเท่ากับ 0.2 ซึ่งเป็นค่าที่ทำให้ประสิทธิภาพของแบบจำลองมีค่าสูงที่สุด

2. เทคนิคเคเนียเรสเนเบอร์ เป็นเทคนิคการจำแนกประเภทของข้อมูล ใช้วิธีการวิเคราะห์จากข้อมูลที่มีคุณสมบัติใกล้เคียงที่สุดจำนวน K ในการหาระยะห่าง ซึ่งในงานวิจัยนี้ได้กำหนดค่า K เพื่อคำนวณระยะห่างเท่ากับ 5 ซึ่งเป็นค่าที่ทำให้ประสิทธิภาพของแบบจำลองมีค่าสูงที่สุด

3. เทคนิคต้นไม้สุ่ม เป็นเทคนิคที่ทำงานคล้ายกับ C4.5 จะเป็นการสุ่มแต่ละปัจจัยมาเป็นโหนดจากต้นไม้หลายๆต้น โดยพารามิเตอร์ของเทคนิคต้นไม้สุ่มเป็นการกำหนดความลึกสุดของต้นไม้ (MaxDepth) งานวิจัยนี้ได้กำหนดความลึกของต้นไม่ตัดสินใจเท่ากับ 10 เพื่อลดโอกาศในการเกิดความซ้ำซ้อนของแบบจำลอง และยังทำให้แบบจำลองมีประสิทธิภาพสูงที่สุด

4. เทคนิคเพอร์เซ็ปตรอนหลายชั้น เป็นเทคนิคที่ใช้ข้อมูลอินพุตและค่าน้ำหนักเพื่อการตัดสินใจของแบบจำลองแต่ละชั้นของข้อมูล (Hidden Layers) พารามิเตอร์คือเลเยอร์ที่ซ่อนอยู่ของโครงข่าย

ประสาทเทียม ในงานวิจัยนี้ได้กำหนดเลเยอร์ที่ซ่อนอยู่ของโครงข่ายประสาทเทียมเท่ากับ 5 ซึ่งเป็นค่าที่ทำให้ประสิทธิภาพของแบบจำลองมีค่าสูงที่สุด

5. เทคนิคลอจิสติกถดถอยแบบต้นไม้ เป็นเทคนิคที่นำฟังก์ชันการถดถอยที่ใบเพื่อเชื่อมโยงทุกโหนดแต่ละปัจจัย

3.4 ขั้นตอนการวัดประสิทธิภาพ

การวัดประสิทธิภาพของแบบจำลองโดยใช้หลักการ 10-fold cross validation ในการพยากรณ์ข้อมูลออกเป็นสองชุด ประกอบด้วยชุดข้อมูลฝึกสอน และชุดข้อมูลทดสอบ โดยจะทำการทดสอบทั้งหมด 10 รอบ เพื่อหาค่าความไว (Sensitivity) คือค่าที่แบบจำลองการรักษาโรคเบาหวานและโรคความดันโลหิตสูงสามารถพยากรณ์ผู้ป่วยที่มีผลการรักษาปกติได้ถูกต้อง ต่อผู้ป่วยทั้งหมด ค่าความจำเพาะ (Specificity) คือค่าที่แบบจำลองการรักษาโรคเบาหวานและโรคความดันโลหิตสูงสามารถพยากรณ์ผู้ป่วยที่มีผลการรักษาได้ถูกต้อง ต่อผู้ป่วยที่มีผลการรักษาผิดปกติทั้งหมด และค่าความถูกต้อง (Accuracy) คือค่าที่แบบจำลองการรักษาโรคเบาหวานและโรคความดันโลหิตสูงสามารถพยากรณ์ผู้ป่วยที่มีผลการรักษาปกติ และผลการรักษาผิดปกติได้ถูกต้อง ต่อผู้ป่วยทั้งหมด

3.5 เครื่องมือที่ใช้ในงานวิจัย

ในงานวิจัยนี้ได้นำเครื่องมือต่างๆ มาใช้ในการค้นหาข้อมูล จัดรูปแบบข้อมูลเพื่อที่จะนำไปวิเคราะห์ข้อมูลและประมวลผลชุดคำสั่ง ประกอบด้วย

1. ชุดคำสั่งภาษา SQL เพื่อใช้ดึงข้อมูลจากฐานข้อมูลโรงพยาบาลทุ่งเขาหลวง
2. โปรแกรม Excel ใช้ในการจัดรูปแบบข้อมูลเพื่อให้สามารถที่จะนำไปวิเคราะห์
3. ชุดโปรแกรม PHP เพื่อใช้ในการกำหนดคลาสให้กับชุดข้อมูล
4. โปรแกรม WAKA ใช้ในการวิเคราะห์ข้อมูลในการพยากรณ์
5. เครื่องคอมพิวเตอร์ ใช้ในการประมวลผลข้อมูลชุดคำสั่ง และโปรแกรมที่ใช้ในการวิเคราะห์

บทที่ 4
ผลการวิจัย

ผลการดำเนินการวิจัยประกอบด้วย ผลการคัดเลือกตัวแปรที่สามารถนำมาพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูงด้วยหลักการ Chi-Square test, Gain Ratio และ Information Grin และเปรียบเทียบประสิทธิภาพของแบบจำลองการพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูง ด้วยเทคนิค Decision Tree (DT C4.5), K-Nearest Neighbors (K-NN), Random Tree (RT), Multi-layer perceptron (MLP) และ Logistic Model Trees (LMT) โดยการทดสอบใช้หลักการ 10-fold cross validation ในการแบ่งข้อมูลออกเป็น ชุดฝึกสอนและชุดทดสอบโดยใช้ชุดฝึกในการสร้างแบบจำลองและชุดทดสอบนำมาทดสอบแบบจำลองด้วยค่าความไว และค่าจำเพาะ และค่าความถูกต้อง

4.1 ผลการคัดเลือกตัวแปรที่สามารถนำมาพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูง

ผลการคัดเลือกตัวแปรหลังจากใช้หลักการการคัดเลือกตัวแปรด้วยวิธี Chi-Square test, Gain Ratio และ Information Gain โดยใช้การทดสอบความสัมพันธ์ระหว่างปัจจัยของชุดข้อมูลโรคเบาหวานและชุดข้อมูลโรคความดันโลหิตสูงโดยคัดเลือกเฉพาะปัจจัยที่มีค่าความสัมพันธ์มากกว่าศูนย์สามารถแสดงปัจจัยที่คัดเลือก และค่าความสัมพันธ์ได้ดังตารางที่ 4.1 และตารางที่ 4.2 ตามลำดับ

ตารางที่ 4.1 ผลการคัดเลือกตัวแปรผู้ป่วยโรคเบาหวาน

ระดับความสัมพันธ์	วิธีการคัดเลือกปัจจัย					
	Chi-Square test		Gain Ratio		Information Gain	
	Select	ค่าความสัมพันธ์	Select	ค่าความสัมพันธ์	Select	ค่าความสัมพันธ์
1	Age	36.900	Bmi	0.023	Age	0.022
2	Occ	28.639	Age	0.022	Occ	0.021
3	Bmi	23.717	Pulse	0.019	Bmi	0.015
4	Pulse	21.340	Dm3	0.018	Pulse	0.013
5	Dm2	16.619	Occ	0.014	Edu	0.010
6	Edu	13.938	Marry	0.013	Dm2	0.010
7	Marry	13.558	Dm2	0.010	Marry	0.008
8	Dm3	11.352	Edu	0.004	Dm3	0.007
9	Dm1	4.065	Dm1	0.003	Dm1	0.002
10	Sex	1.111	Sex	0.001	Sex	0.001

จากตารางที่ 4.1 แสดงผลการคัดเลือกตัวแปรหลังจากใช้วิธีการคัดเลือกปัจจัย จากชุดข้อมูลการรักษาโรคเบาหวานโดยในทุกหลักการจะเลือกตัวแปรที่มีค่าความสัมพันธ์มากกว่าศูนย์ ส่วนตัวแปรที่มีความสัมพันธ์น้อยกว่าศูนย์ หรือที่ไม่มีค่าจะถูกตัดออก ตัวแปรที่เหลือจะถูกนำไปสร้างแบบจำลอง จากทั้งหมด 14 ตัวแปร หลังการคัดเลือกเหลือเพียง 10 ตัวแปรเท่ากันทั้ง 3 หลักการและเหมือนกันทุกตัวแปร แต่จะแตกต่างกันที่การจัดระดับความสัมพันธ์ระหว่างตัวแปรในการคัดเลือกตัวแปร เช่น วิธี Chi-Square test สามารถจัดระดับความสัมพันธ์ของตัวแปรในการคัดเลือกได้แก่ Age, Occ, Bmi, Pulse, Dm2, Edu, Marry, Dm3, Dm1, Sex ตามลำดับ จะแตกต่างกับวิธี Gain Ratio ที่จัดระดับความสัมพันธ์ของตัวแปรในการคัดเลือกได้แก่ Bmi, Age, Pluse, Dm3, Occ, Marry, Dm2, Edu, Dm1, Sex ตามลำดับ ส่วนวิธี Information Grin สามารถจัดระดับ

ความสัมพันธ์ของตัวแปรในการคัดเลือกตัวแปร ได้แก่ Age, Occ, Bmi, Pulse, Edu, Dm2, Marry, Dm3, Dm1, Sex ตามลำดับ

ตารางที่ 4.2 ผลการคัดเลือกตัวแปรผู้ป่วยโรคความดันโลหิตสูง

ระดับความสัมพันธ์	วิธีการคัดเลือกปัจจัย					
	Chi-Square test		Gain Ratio		Information Gain	
	Select	ค่าความสัมพันธ์	Select	ค่าความสัมพันธ์	Select	ค่าความสัมพันธ์
1	Bps	88.320	Bps	0.052	Bps	0.131
2	Ht2	35.599	Occ	0.009	Occ	0.011
3	Marry	4.003	Ht3	0.003	Marry	0.002
4	Ht3	3.849	Marry	0.003	Ht3	0.001
5	Edu	3.849	Ht2	0.001	Ht2	0.0004
6	Ht1	1.280	Ht1	0.0003	Ht1	0.0003
7	Sex	1.221	Edu	0.0001	Sex	0.0002

จากตารางที่ 4.2 แสดงผลการคัดเลือกตัวแปรหลังจากใช้หลักการการคัดเลือกปัจจัย จากชุดข้อมูลการรักษาโรคความดันโลหิตสูงโดยในทุกหลักการจะเลือกตัวแปรที่มีค่าความสัมพันธ์มากกว่าศูนย์ ส่วนตัวแปรที่มีความสัมพันธ์น้อยกว่าศูนย์หรือที่ไม่มีค่าจะถูกตัดออก ตัวแปรที่เหลือจะถูกนำไปสร้างแบบจำลองจากทั้งหมด 13 ตัวแปร หลังการคัดเลือกเหลือเพียง 7 ตัวแปรเท่ากันทั้ง 3 วิธี แต่ตัวแปรที่ถูกคัดเลือกมีความแตกต่างกัน คือ วิธี Chi-Square test ไม่เลือก Occ วิธี Gain Ratio ไม่เลือก Sex ส่วนวิธี Information Gain ไม่เลือก Edu

เมื่อได้ข้อมูลที่คัดเลือกตัวแปรด้วยวิธี Chi-Square test, Gain Ratio และ Information Grin แล้วนำข้อมูลที่ถูกคัดเลือกไปสร้างแบบจำลองเพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองในการพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูง

4.2 ผลการเปรียบเทียบประสิทธิภาพของแบบจำลองพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูง

4.2.1 ผลค่าความไวของแบบจำลอง

ค่าความไวของแบบจำลองพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูงก่อนและหลักการคัดเลือกตัวแปรด้วยวิธี Chi-Square test, Gain Ratio และ Information Grin แล้วนำมาสร้างแบบจำลองพยากรณ์ด้วยเทคนิค Decision Tree (DT C4.5), K-Nearest Neighbors (K-NN), Random Tree (RT), Multi-layer perceptron (MLP) และ Logistic Model Trees (LMT) ได้ผลดังตารางที่ 4.3 และตารางที่ 4.4 ตามลำดับ

ตารางที่ 4.3 ค่าความไวของแบบจำลองผู้ป่วยโรคเบาหวานก่อนและหลังการคัดเลือกตัวแปร

แบบจำลอง	ก่อนการคัดเลือกตัวแปร	การคัดเลือกตัวแปร		
		Chi-Square test	Gain Ratio	Information Grin
DT C4.5	87.57	93.20(↑5.63)	94.49(↑6.92)	94.49(↑6.92)
K-NN	70.34	82.30(↑11.96)	82.30(↑11.96)	82.30(↑11.96)
RT	71.04	81.36(↑10.32)	88.98(↑17.94)	73.62(↑2.58)
MLP	82.06	88.04(↑5.98)	88.04(↑5.98)	88.04(↑5.98)
LMT	94.84	95.08(↑0.24)	95.08(↑0.24)	95.08(↑0.24)

จากตารางที่ 4.3 แสดงค่าความไวของแบบจำลองผู้ป่วยโรคเบาหวานก่อนและหลังการคัดเลือกตัวแปรและผลต่างในการคัดเลือก ด้วยวิธี Chi-Square test, Gain Ratio และ Information Grin โดยใช้เทคนิค Decision Tree (DT C4.5), K-Nearest Neighbors (K-NN), Random Tree (RT), Multi-layer perceptron (MLP) และ Logistic Model Trees (LMT) ในการสร้างแบบจำลองการพยากรณ์การรักษาโรค ผลปรากฏว่า ก่อนการคัดเลือกตัวแปรแบบจำลองที่สร้างด้วยเทคนิค Logistic Model Trees (LMT) ให้ค่าความไวสูงสุดที่ร้อยละ 94.84 แบบจำลองที่สร้างด้วยเทคนิค K-Nearest Neighbors (K-NN) ให้ค่าความไวต่ำสุดที่ร้อยละ 70.34 หลังคัดเลือกตัวแปร Gain Ratio ช่วยให้แบบจำลองที่สร้างด้วยเทคนิค Random Tree (RT) มีผลต่างสูงสุดถึงร้อยละ 17.94 แต่อย่างไรก็ตามทุกวิธีการคัดเลือกตัวแปรทำให้แบบจำลองที่สร้างด้วยเทคนิค Logistic Model Trees

(LMT) ให้ค่าความไวสูงสุดที่ร้อยละ 95.08 แต่การคัดเลือกตัวแปรด้วยวิธี Gain Ratio และ Information Grin ทำให้แบบจำลองที่สร้างด้วยเทคนิค DT C4.5 มีค่าความไวรองลงมาสูงถึงร้อยละ 94.49

ตารางที่ 4.4 ค่าความไวของแบบจำลองผู้ป่วยโรคความดันโลหิตสูงก่อนและหลังการคัดเลือกตัวแปร

แบบจำลอง	ก่อนการคัดเลือกตัวแปร	การคัดเลือกตัวแปร		
		Chi-Square test	Gain Ratio	Information Grin
DT 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(0.00)
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)

จากตารางที่ 4.4 แสดงค่าความไวของแบบจำลองผู้ป่วยโรคความดันโลหิตสูงก่อนและหลังการคัดเลือกตัวแปรและผลต่างในการคัดเลือก ด้วยวิธี Chi-Square test, Gain Ratio และ Information Grin โดยใช้เทคนิค Decision Tree (DT C4.5), K-Nearest Neighbors (K-NN), Random Tree (RT), Multi-layer perceptron (MLP) และ Logistic Model Trees (LMT) ในการสร้างแบบจำลองการพยากรณ์การรักษาโรค ผลปรากฏว่า ก่อนการคัดเลือกตัวแปรแบบจำลองที่สร้างด้วยเทคนิค Logistic Model Trees (LMT) ให้ค่าความไวสูงสุดที่ร้อยละ 93.92 แบบจำลองที่สร้างด้วยเทคนิค Random Tree (RT) ให้ค่าความไวต่ำสุดที่ร้อยละ 88.97 หลังคัดเลือกตัวแปร Information Grin ช่วยให้แบบจำลองที่สร้างด้วยเทคนิค Multi-layer perceptron (MLP) มีผลต่างสูงสุดถึงร้อยละ 2.26 แต่อย่างไรก็ตามการคัดเลือกตัวแปรด้วยวิธี Chi-Square test ทำให้แบบจำลองที่สร้างด้วยเทคนิค Decision Tree (DT C4.5) มีค่าความไวสูงที่สุดร้อยละ 93.89 แต่การคัดเลือกตัวแปรด้วย Information Grin ทำให้แบบจำลองที่สร้างด้วยเทคนิค Multi-layer perceptron (MLP) มีค่าความไวรองลงมาสูงถึงร้อยละ 93.86

4.2.2 แสดงค่าความจำเพาะของแบบจำลอง

ค่าความจำเพาะของแบบจำลองพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูง ก่อนและหลักการคัดเลือกตัวแปรด้วยวิธี Chi-Square test, Gain Ratio และ Information Grin แล้วนำมาสร้างแบบจำลองพยากรณ์ด้วยเทคนิค Decision Tree (DT C4.5), K-Nearest Neighbors (K-NN), Random Tree (RT), Multi-layer perceptron (MLP) และ Logistic Model Trees (LMT) ได้ผลดังตารางที่ 4.5 และตารางที่ 4.6 ตามลำดับ

ตารางที่ 4.5 ค่าความจำเพาะของแบบจำลองผู้ป่วยโรคเบาหวานก่อนและหลังการคัดเลือกตัวแปร

แบบจำลอง	ก่อนการคัดเลือกตัวแปร	การคัดเลือกตัวแปร		
		Chi-Square test	Gain Ratio	Information Grin
DT C4.5	45.36	40.51(↓4.85)	40.51(↓4.85)	40.51(↓4.85)
K-NN	37.22	38.87(↑1.65)	38.87(↑1.65)	38.87(↑1.65)
RT	37.47	38.16(↑0.69)	38.16(↑0.69)	41.10(↑3.63)
MLP	44.77	41.38(↓3.39)	41.38(↓3.39)	41.38(↓3.39)
LMT	54.17	54.35(↑0.18)	54.35(↑0.18)	54.35(↑0.18)

จากตารางที่ 4.5 แสดงค่าความจำเพาะของแบบจำลองผู้ป่วยโรคเบาหวานก่อนและหลังการคัดเลือกตัวแปรและผลต่างในการคัดเลือก ด้วยวิธี Chi-Square test, Gain Ratio และ Information Grin โดยใช้เทคนิค Decision Tree (DT C4.5), K-Nearest Neighbors (K-NN), Random Tree (RT), Multi-layer perceptron (MLP) และ Logistic Model Trees (LMT) ในการสร้างแบบจำลองการพยากรณ์การรักษาโรค ผลปรากฏว่า ก่อนการคัดเลือกตัวแปร แบบจำลองที่สร้างด้วยเทคนิค Logistic Model Trees (LMT) ให้ค่าความจำเพาะสูงสุดที่ร้อยละ 54.17 แบบจำลองที่สร้างด้วยเทคนิค K-Nearest Neighbors (K-NN) ให้ค่าความจำเพาะต่ำสุดที่ร้อยละ 37.22 หลังคัดเลือกตัวแปร Information Grin ช่วยให้แบบจำลองที่สร้างด้วยเทคนิค Random Tree (RT) มีผลต่างสูงสุดถึงร้อยละ 3.36 แต่อย่างไรก็ตามทุกวิธีการคัดเลือกตัวแปรทำให้แบบจำลองที่สร้างด้วยเทคนิค Logistic Model Trees (LMT) มีค่าความจำเพาะสูงที่สุดร้อยละ 54.35 แต่การคัดเลือกตัวแปรด้วย

วิธี Gain Ratio และ Information Grin ทำให้แบบจำลองที่สร้างด้วยเทคนิค Multi-layer perceptron (MLP) มีค่าความจำเพาะรองลงมาสูงถึงร้อยละ 41.38

ตารางที่ 4.6 ค่าความจำเพาะของแบบจำลองผู้ป่วยโรคความดันโลหิตสูงก่อนและหลังการคัดเลือกตัวแปร

แบบจำลอง	ก่อนการคัดเลือกตัวแปร	การคัดเลือกตัวแปร		
		Chi-Square test	Gain Ratio	Information Grin
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)

จากตารางที่ 4.6 แสดงค่าความจำเพาะของแบบจำลองผู้ป่วยโรคความดันโลหิตสูงก่อนและหลังการคัดเลือกตัวแปรและผลต่างในการคัดเลือก ด้วยวิธี Chi-Square test, Gain Ratio และ Information Grin โดยใช้เทคนิค Decision Tree (DT C4.5), K-Nearest Neighbors (K-NN), Random Tree (RT), Multi-layer perceptron (MLP) และ Logistic Model Trees (LMT) ในการสร้างแบบจำลองการพยากรณ์การรักษาโรค ผลปรากฏว่า ก่อนการคัดเลือกตัวแปร แบบจำลองที่สร้างด้วยเทคนิค Logistic Model Trees (LMT) ให้ค่าความจำเพาะสูงสุดที่ร้อยละ 62.95 แบบจำลองที่สร้างด้วยเทคนิค K-Nearest Neighbors (K-NN) ให้ค่าความจำเพาะต่ำสุดที่ร้อยละ 48.75 หลังคัดเลือกตัวแปร Gain Ratio ช่วยให้แบบจำลองที่สร้างด้วยเทคนิค K-Nearest Neighbors (K-NN) มีผลต่างสูงสุดถึงร้อยละ 15.35 แต่อย่างไรก็การคัดเลือกตัวแปรด้วยวิธี Chi-Square test ทำให้แบบจำลองที่สร้างด้วยเทคนิค Decision Tree (DT C4.5) มีค่าความจำเพาะสูงที่สุดร้อยละ 62.88 แต่แบบจำลองที่สร้างด้วยเทคนิค Logistic Model Trees (LMT) มีค่าความไวรองลงมาสูงถึงร้อยละ 62.35

4.2.3 แสดงค่าความถูกต้องของแบบจำลอง

ค่าความถูกต้องของแบบจำลองพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูง หลักการคัดเลือกตัวแปรด้วยวิธี Chi-Square test, Gain Ratio และ Information Grin แล้วนำมาสร้างแบบจำลองพยากรณ์ด้วยเทคนิค Decision Tree (DT C4.5), K-Nearest Neighbors (K-NN), Random Tree (RT), Multi-layer perceptron (MLP) และ Logistic Model Trees (LMT) ได้ผลดังตารางที่ 4.7 และตารางที่ 4.8 ตามลำดับ

ตารางที่ 4.7 ค่าความถูกต้องของแบบจำลองผู้ป่วยโรคเบาหวานก่อนและหลังการคัดเลือกตัวแปร

แบบจำลอง	ก่อนการคัดเลือกตัวแปร	การคัดเลือกตัวแปร		
		Chi-Square test	Gain Ratio	Information Grin
DT C4.5	67.89	68.13(↑0.24)	68.13(↑0.24)	68.13(↑0.24)
K-NN	60.98	64.88(↑3.90)	64.88(↑3.90)	64.88(↑3.90)
RT	61.30	66.96(↑5.56)	66.42(↑5.12)	63.82(↑2.52)
MLP	66.99	66.91(↓0.08)	66.91(↓0.08)	66.91(↓0.08)
LMT	70.00	70.00(0.00)	70.00(0.00)	70.00(0.00)

จากตารางที่ 4.7 แสดงค่าความถูกต้องของแบบจำลองผู้ป่วยโรคเบาหวานก่อนและหลังการคัดเลือกตัวแปรและผลต่างในการคัดเลือก ด้วยวิธี Chi-Square test, Gain Ratio และ Information Grin โดยใช้เทคนิค Decision Tree (DT C4.5), K-Nearest Neighbors (K-NN), Random Tree (RT), Multi-layer perceptron (MLP) และ Logistic Model Trees (LMT) ในการสร้างแบบจำลองการพยากรณ์การรักษาโรค ผลปรากฏว่า ก่อนการคัดเลือกตัวแปร แบบจำลองที่สร้างด้วยเทคนิค Logistic Model Trees (LMT) ให้ค่าความถูกต้องสูงสุดที่ร้อยละ 70.00 แบบจำลองที่สร้างด้วยเทคนิค K-Nearest Neighbors (K-NN) ให้ค่าความถูกต้องต่ำสุดที่ร้อยละ 60.98 หลังคัดเลือกตัวแปรด้วยวิธี Chi-Square test ช่วยให้แบบจำลองที่สร้างขึ้นด้วย Random Tree (RT) มีผลต่างสูงถึงร้อยละ 5.56 แต่อย่างไรก็ตามทุกวิธีการคัดเลือกตัวแปรทำให้แบบจำลองที่สร้างด้วยเทคนิค Logistic Model Trees (LMT) มีค่าความถูกต้องสูงที่สุดร้อยละ 70.00 แต่การคัดเลือกตัวแปรทุกวิธีการ ทำให้

แบบจำลองที่สร้างด้วยเทคนิค Decision Tree (DT C4.5) มีค่าความถูกต้องรองลงมาสูงถึงร้อยละ 68.13

ตารางที่ 4.8 ค่าความถูกต้องของแบบจำลองผู้ป่วยโรคความดันโลหิตสูงก่อนและหลังการคัดเลือกตัวแปร

แบบจำลอง	ก่อนการคัดเลือกตัวแปร	การคัดเลือกตัวแปร		
		Chi-Square test	Gain Ratio	Information Grin
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)

จากตารางที่ 4.8 แสดงค่าความถูกต้องของแบบจำลองผู้ป่วยโรคความดันโลหิตสูงก่อนและหลังการคัดเลือกตัวแปรและผลต่างในการคัดเลือก ด้วยวิธี Chi-Square test, Gain Ratio และ Information Grin โดยใช้เทคนิค Decision Tree (DT C4.5), K-Nearest Neighbors (K-NN), Random Tree (RT), Multi-layer perceptron (MLP) และ Logistic Model Trees (LMT) ในการสร้างแบบจำลองการพยากรณ์การรักษาโรค ผลปรากฏว่า ก่อนการคัดเลือกตัวแปร แบบจำลองที่สร้างด้วยเทคนิค Logistic Model Trees (LMT) ให้ค่าความถูกต้องสูงสุดที่ร้อยละ 80.24 แบบจำลองที่สร้างด้วยเทคนิค K-Nearest Neighbors (K-NN) ให้ค่าความถูกต้องต่ำสุดที่ร้อยละ 76.06 หลังคัดเลือกตัวแปรด้วยวิธีการ Gain Ratio ช่วยให้แบบจำลองที่สร้างขึ้นด้วย K-Nearest Neighbors (K-NN) มีผลต่างสูงถึงร้อยละ 3.72 แต่อย่างไรก็ตามการคัดเลือกตัวแปรด้วยวิธี Chi-Square test ทำให้แบบจำลองที่สร้างด้วยเทคนิค Decision Tree (DT C4.5) มีค่าความถูกต้องสูงสุดที่ร้อยละ 80.82 แต่การคัดเลือกตัวแปรด้วยวิธี Chi-Square test และ Information Grin ทำให้แบบจำลองที่สร้างด้วยเทคนิคLogistic Model Trees (LMT) มีค่าความถูกต้องรองลงมาสูงถึงร้อยละ 80.36

บทที่ 5
สรุป อภิปรายและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์ เพื่อศึกษาปัจจัยในการคัดเลือกตัวแปรในการสร้างแบบจำลองการพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูง และเปรียบเทียบประสิทธิภาพของแบบจำลองที่สร้างขึ้น ในงานวิจัยนี้ได้รวบรวมข้อมูลการรักษาโรคเบาหวานและโรคความดันโลหิตสูงจากฐานข้อมูลโรงพยาบาลทุ่งเขาหลวง อำเภอทุ่งเขาหลวง จังหวัดร้อยเอ็ด จำนวนทั้งหมด 2 ชุดข้อมูล คัดเลือกปัจจัยด้วยวิธีการ Chi-Square test, Gain Ratio และ Information Grin ด้วยเทคนิค Decision Tree (DT C4.5) เทคนิค K-Nearest Neighbors (K-NN) เทคนิค Random Tree (RT) เทคนิค Multi-layer perceptron (MLP) และเทคนิค Logistic Model Trees (LMT) จากการทดลองสรุปได้ดังนี้

จากการเปรียบเทียบประสิทธิภาพของแบบจำลองการพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูงที่สร้างขึ้น ในชุดข้อมูลโรคเบาหวานก่อนการคัดเลือกตัวแปร เทคนิคที่ดีที่สุดคือ Logistic Model Trees (LMT) ให้ค่าความไวมากที่สุดถึงร้อยละ 94.84 ค่าความจำเพาะสูงถึงร้อยละ 62.95 ค่าความถูกต้องสูงที่สุดร้อยละ 70.00 ส่วนข้อมูลโรคความดันโลหิตสูง ก่อนการคัดเลือกตัวแปร เทคนิคที่ดีที่สุดคือ Decision Tree (DT C4.5) ให้ค่าความไวสูงที่สุดร้อยละ 93.70 ค่าความจำเพาะด้วยเทคนิค Logistic Model Trees (LMT) สูงที่สุดร้อยละ 62.95 และยังให้ค่าความถูกต้องสูงที่สุดร้อยละ 80.24

หลังการคัดเลือกตัวแปร ในชุดข้อมูลโรคเบาหวานโดยใช้หลักการคัดเลือกตัวแปรด้วยวิธี Gain Ratio ร่วมกับเทคนิค Random Tree (RT) สามารถเพิ่มค่าความไวให้กับแบบจำลองมากที่สุดร้อยละ 17.94 ค่าความจำเพาะด้วยวิธี Information Gain ร่วมกับเทคนิค Random Tree (RT) ให้ค่าเพิ่มมากที่สุดร้อยละ 3.63 และค่าความถูกต้องด้วยวิธี Chi-Square test ร่วมกับเทคนิค Random Tree (RT) ให้ค่าเพิ่มมากที่สุดร้อยละ 5.56 แต่ถึงอย่างไรวิธี Chi-Square test ร่วมกันเทคนิค Logistic Model Trees (LMT) ให้ค่าความไวสูงที่สุดร้อยละ 95.08 ให้ค่าสูงขึ้นหลังจากการคัดเลือกตัวแปรร้อยละ 0.24 ค่าความจำเพาะสูงที่สุดร้อยละ 54.35 ให้ค่าสูงขึ้นหลังจากการคัดเลือกตัวแปรร้อยละ 0.18 และค่าความถูกต้องสูงที่สุดร้อยละ 70 ส่วนชุดข้อมูลโรคความดันโลหิตสูงโดยใช้หลักการคัดเลือกตัวแปรด้วยวิธีการ Information Gain ร่วมกับเทคนิค Multi-layer perceptron (MLP) สามารถเพิ่มค่าความไวให้กับแบบจำลองมากที่สุดร้อยละ 2.26 ค่าความจำเพาะด้วยวิธี Gain Ratio ร่วมกับเทคนิค K-Nearest Neighbors (K-NN) ให้ค่าเพิ่มมากที่สุดร้อยละ 15.35 และยังให้ค่าความ

ถูกต้องเพิ่มมากที่สุดร้อยละ 3.72 แต่ถึงอย่างไรวิธี Chi-Square test ร่วมกันเทคนิค Decision Tree (DT C4.5) ให้ค่าความไวสูงที่สุดร้อยละ 93.89 ให้ค่าสูงขึ้นหลังจากการคัดเลือกตัวแปรร้อยละ 0.19 ค่าความจำเพาะสูงที่สุดร้อยละ 62.88 ให้ค่าสูงขึ้นหลังจากการคัดเลือกตัวแปรร้อยละ 1.75 และค่าความถูกต้องสูงที่สุดร้อยละ 80.82 ให้ค่าสูงขึ้นหลังจากการคัดเลือกตัวแปรร้อยละ 0.12

5.2 อภิปรายผล

จากการศึกษาปัจจัยในการคัดเลือกตัวแปรในการสร้างแบบจำลองพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูง และเปรียบเทียบประสิทธิภาพของแบบจำลองการพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูง ด้วยเทคนิคเหมืองข้อมูล ผลการทดลองพบว่า ในชุดข้อมูลโรคเบาหวานการคัดเลือกตัวแปรด้วยวิธีการ Chi-Square test ร่วมกับเทคนิค Logistic Model Trees (LMT) ให้ประสิทธิภาพดีที่สุด เมื่อเทียบกับการคัดเลือกตัวแปรด้วยวิธี Gain Ratio และ Information Grin ร่วมกับเทคนิค Tree (DT C4.5) เทคนิค K-Nearest Neighbors (K-NN) เทคนิค Random Tree (RT) และเทคนิค Multi-layer perceptron (MLP) ในชุดข้อมูลโรคความดันโลหิตสูงการคัดเลือกตัวแปรด้วยวิธีการ Chi-Square test ร่วมกับเทคนิค Decision Tree (DT C4.5) ให้ประสิทธิภาพดีที่สุด เมื่อเทียบกับการคัดเลือกตัวแปรด้วยวิธี Gain Ratio และ Information Grin ร่วมกับเทคนิค K-Nearest Neighbors (K-NN) เทคนิค Random Tree (RT) เทคนิค Multi-layer perceptron (MLP) และเทคนิค Logistic Model Trees (LMT) สรุปได้ว่าในงานวิจัยนี้การคัดเลือกตัวแปรด้วยวิธี Chi-Square test เหมาะสมที่สุดในการนำมาสร้างแบบจำลองพยากรณ์การรักษาโรคเบาหวานร่วมกับเทคนิค Logistic Model Trees (LMT) และโรคความดันโลหิตสูงร่วมกับเทคนิค Decision Tree (DT C4.5) ซึ่งสอดคล้องกับงานวิจัยของ อุกฤษฏ์ ศรีสุข และจารี ทองคำ [9] ได้ทำการเปรียบเทียบประสิทธิภาพของแบบจำลองของเทคนิคเหมืองข้อมูลสำหรับพยากรณ์การเกิดโรค พบว่า การคัดเลือกตัวแปรด้วยวิธี Chi-Square test ร่วมกับเทคนิค Decision Tree (DT C4.5) ให้ค่าความถูกต้องสูงที่สุด ซึ่งมีความเหมาะสมในการสร้างแบบจำลองในการพยากรณ์โรคมากที่สุด

5.3 ข้อเสนอแนะ

ในงานวิจัยนี้เป็นการสร้างแบบจำลองในการพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูง หากนำไปใช้กับชุดข้อมูลอื่น อาจต้องเพิ่มเทคนิคเหมืองข้อมูลเทคนิคอื่นๆเข้ามาเปรียบเทียบหรือวิธีการคัดเลือกตัวแปรในการพยากรณ์ หรืออาจเพิ่มข้อมูลการรักษาโรคอื่นๆเข้ามาร่วมด้วย เพื่อให้ได้ผลการพยากรณ์ที่มีประสิทธิภาพมากยิ่งขึ้นไป

บรรณานุกรม

เอกสารอ้างอิง

[1] นพ.สมเกียรติโพธิสัตย์. (16/07/2564). โรคเบาหวาน (*Diabetes Mellitus)*. Available: https://www.hiso.or.th

[2] น. ไชยอัชนรัตน์. (19/07/2564). เบาหวานโรคไกล้ตัวที่ควรรู้. Available: https://www.sikarin.com/doctor-articles

[3] โรงพยาบาลกรุงเทพภูเก็ต. (2561, 17/07/2564). โรคเบาหวาน. Available: https://www.phukethospital.com

[4] โรงพยาบาลศิริราช ปิยมหาราชการุณย์. (2563, 17/07/2564). คุณเป็นโรคความดันโลหิตสูงหรือเปล่า. Available: https://www.siphhospital.com

[5] กรมวิทยาศาสตร์การแพทย์. (17/07/2564). โรคความดันโลหิตสูงคืออะไร. Available: https://xn--12c8b3afcz5g8i.com/

[6] ณัฐธิวรรณ พันธ์มุง, อลิสรา อยู่เลิศลบ, และ สราญรัตน์ ลัทธิ. (2562). ประเด็นสารรณรงค์วันความดันโลหิตสูงโลก ปี2562. Available: http://www.thaincd.com

[7] ขนิษฐา พิศฉลาด และ ฉันทนา แรงสินห์, "การพัฒนารูปแบบการป้องกันโรคเบาหวานชนิดที่ 2 สำหรับกลุ่มเสี่ยงโรคเบาหวานในชุมชน จังหวัดเชียงราย," 132-146, vol. 28, no. 3, 2559.

[8] เพชรรัตน์ ม่วงน้อย, จักรพันธ์ พลาผล, และ ภรัณยา ปาลวิสุทธ, "ตัวแบบประเมินภาวะความเสี่ยงการเป็นโรคซึมเศร้าของนักศึกษาด้วยเทคนิคเหมืองข้อมูล," การประยุกต์ใช้เทคโนโลยีสารสนเทศ, vol. 7, no. 1, 2564.

[9] อุกฤษฎ์ ศรีสุข และ จารี ทองคำ, "การเปรียบเทียบประสิทธิภาพของเทคนิคเหมืองข้อมูลสำหรับพยากรณ์การเกิดโรค," ภาควิชาเทคโนโลยีสารสนเทศ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม, vol. 40, no. 2, 2564.

[10] S. A. Aboalnaser and H. R. Almohammadi, "Comprehensive Study of Diabetes Miletus Prediction Using Different Classification Algorithms," in 2019 12*th International Conference on Developments in eSystems Engineering (DeSE)*, 2019, pp. 128-133.

[11] อัจราภรณ์ สุขเพิ่ม และคณะ "แบบจำลองการวินิจฉัยอัตโนมัติสำหรับความเสี่ยงต่อการเกิดลิ่มเลือดอุดตันในหลอดเลือดดำตามอาการโดยอาศัยการเรียนรู้ของเครื่อง," *Proceeding of the Data science conference,* 2021.

[12] บรรจบ ดลกุล, และ จารี ทองคำ "การสร้างแบบจำลองเพื่อพยากรณ์การเกิดแผลที่เท้าของผู้ป่วยโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูล," *j Sci Technol MSU,* vol. 33, no. 6, 2014.

[13] Sujan Ray, Khaldoon Alshouiliy, Anupam Roy , Ali AlGhamdi, and D. P. Agrawal, "Chi-Squared Based Feature Selection for Stroke Prediction using AzureML," *Intermountain Engineering, Technology and Computing (IETC),* 2020.

[14] S. sittichat, "Stude of Educational Attributes Using Data Mining Technique," *Rajamangala University of Tachnology Lanna Chiang Mai,* vol. 13, no. 2, 2017.

[15] อัชฌาพร กว้างสวาสดิ์ และคณะ "ระบบทำนายระดับความเครียด ด้วยเทคนิคต้นไม้การตัดสินใจ," *Rattanakosin Journal of Scienceand Technology:,* vol. 1, no. 2, 2562.

[16] วัฒนพัชร์ เก้าลิ้ม, "การค้นหาคำสำคัญในเอกสารภาษาไทยโดยใช้เทคนิคการค้นหารูปแบบความสัมพันธ์ในเหมืองข้อมูล," บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร, 2557.

[17] รุจิรา ธรรมสมบัต, "ระบบสนับสนุนการตัดสินใจในการเลือกใช้แพคเกจอินเทอร์เน็ตมือถือโดยใช้ต้นไม้ตัดสินใจ" วิทยาลัยราชพฤกษ์, 2554.

[18] กรสิริณัฐ โรจนวรรณ์ และ วิยุดา เพชรจิรโชติกุล, "การเลือกคุณลักษณะที่เหมาะสมสำหรับการแทนค่าข้อมูลสูญหายของข้อมูลอนุกรมเวลาด้วยเทคนิคเหมืองข้อมูล," คณะวิศวกรรมศาสตร์ มหาวิทยาลัยนราธิวาสราชนครินทร์, vol. 13, no. 2, 2564.

[19] K. Alpan and G. S. İlgi, "Classification of Diabetes Dataset with Data Mining Techniques by Using WEKA Approach," in 2020 4*th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 2020, pp. 1-7.

[20] นที ไทยธรรม และ ภาสกร สุวรรณโท. (2562, 7/8/2564). *KNN* หรือ *K-Nearest Neighbors* คืออะไร. Available: https://www.glurgeek.com

[21] ธีรรัศมี, "การหาค่า เค ที่เหมาะสมในการจำแนกแบบเคเนียร์เรสเนเบอร์กับข้อมูลทางการแพทย์," มหาวทิยาลัยเทคโนโลยีสุรนารี, 2558.

[22] พัฒนพงษ์ ดลรัตน์ และ จารี ทองคำ, "การเปรียบเทียบประสิทธิภาพของแบบจำลองในการพยากรณ์ความสำเร็จการศึกษาของนักเรียนระดับประกาศนียบัตรวิชาชีพ," คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม, vol. 37, no. 3, 2560.

[23] กริช สมกันธา, วิไลพร กุลตังวัฒนา, และ วรวิทย์ กุลตังวัฒนา, "การพัฒนาระบบประเมินบุคลากรโดยใช้โครงข่ายประสาทเทียม," vol. 9, no. 1, 2556.

[24] Cai Jun, Zhai Fan, and Feng Shan, "Building up Multi-Layered Perceptrons as Classifier System for Decision Support," *Huazhong University of Science and Technology,* vol. 6, no. 2, 1995.

[25] N. LANDWEHR, "Logistic Model Trees," 2005 *Springer Science + Business Media, Inc. Manufactured in The Netherlands,* vol. 59, no. 161-205, 2005.

[26] นงเยาว์ ในอรุณ, "การเปรียบเทียบประสิทธิภาพของแบบจำ ลองการทำ นายความเสี่ยงโรคหัวใจและหลอดเลือดโดยใช้อัลกอริทึมเหมืองข้อมูล," คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏพระนครศรีอยุธยา, vol. 40, no. 2, 2564.

[27] เอพร โมพี, นิธิศ เสาแก้ว, และ บุษยมาศ เหมณี, "การศึกษาสภาวะเสี่ยงโรคของผู้สูงอายุด้วยเทคนิคเหมืองข้อมูล," วารสารนาคบุตรปริทรรศน์ มหาวิทยาลัยราชภัฏนครศรีธรรมราช, vol. 11, no. 3, 2562.

[28] S. Bamal, M. Gupta, N. Sewal, and A. Sharma, "Performance Comparison of Classification Models for Diabetes Prediction," in 2020 *9th International Conference System Modeling and Advancement in Research Trends (SMART)*, 2020, pp. 45-51.

[29] Kuannabha Pimcharee and O. Surinta, "Data Mining Approaches in Personal Loan Approval," *Department of Information Technology, Faculty of Informatics, Mahasarakham University, Mahasarakham, Thailand,* vol. 8, no. 1, 2022.

[30] Kronsirinut Rothjanawan and W. Phetjirachotkul, "Feature Selection Methods for Imputation Missing Valuesof Time Series Data using Data Mining," *Princess of Naradhiwas University Journal* vol. 13, no. 2, 2021.

[31] Patharawut Saengsiri, Sageemas Na Wichian, and P. Meesad, "Classification of Leukemia Data Using Ranking and SupportVector Machine," *KKU Res J,* vol. 10, no. 2, 2010.

[32] พรรณธิภา เพชรบุญมี, ดวงกมล โพธิ์นาค, และ มนต์ชัย เทียนทอง, "การพยากรณ์รูปแบบการเรียนรู้ตามประสบการณ์ของเดวิด โคล์ป โดยใช้กฎการจำแนกเทคนิคต้นไม้ตัดสินใจ," วารสารวิทยาศาสตร์และเทคโนโลยี, vol. 21, no. 6, 2556.

ภาคผนวก

คณะกรรมการจริยธรรมการวิจัยในคน มหาวิทยาลัยมหาสารคาม

เอกสารรับรองโครงการวิจัย

เลขที่การรับรอง : 035-033/2565

ชื่อโครงการวิจัย (ภาษาไทย) แบบจำลองสำหรับพยากรณ์การรักษาโรคเบาหวานและโรคความดันโลหิตสูงโดยเทคนิคเหมืองข้อมูล

ชื่อโครงการวิจัย (ภาษาอังกฤษ) PREDCTIOIN MODELS FOR DIABETES AND HIGH BLOOD PRESSURE TREATMENTS USING DATA MINING TECHNIQUES

ผู้วิจัย : นายธงไชย พ้องเสียง
หน่วยงานที่รับผิดชอบ : คณะวิทยาการสารสนเทศ
สถานที่ทำการวิจัย : คณะวิทยาการสารสนเทศ

ประเภทการพิจารณาแบบ : แบบยกเว้น

วันที่รับรอง : 31 มกราคม 2565 **วันหมดอายุ** : 30 มกราคม 2566

ข้อเสนอการวิจัยนี้ ได้รับการพิจารณาและให้ความเห็นชอบจากคณะกรรมการจริยธรรมการวิจัยในคน มหาวิทยาลัยมหาสารคามแล้ว และอนุมัติในด้านจริยธรรมให้ดำเนินการศึกษาวิจัยเรื่องข้างต้นได้ บนพื้นฐานของโครงร่างงานวิจัยที่คณะกรรมการฯ ได้รับและพิจารณา เมื่อเสร็จสิ้นโครงการแล้วให้ผู้วิจัยส่งแบบฟอร์มการปิดโครงการและรายงานผลการดำเนินงานมายังคณะกรรมการจริยธรรมการวิจัยในคน มหาวิทยาลัยมหาสารคาม หรือหากมีการเปลี่ยนแปลงใดๆ ในโครงการวิจัย ผู้วิจัยจักต้องยื่นขอรับการพิจารณาใหม่

ภาธร สว่างจิตร

(ผู้ช่วยศาสตราจารย์ เภสัชกรหญิงราตรี สว่างจิตร)
ประธานคณะกรรมการจริยธรรมการวิจัยในคน
มหาวิทยาลัยมหาสารคาม

ทั้งนี้ การรับรองนี้มีเงื่อนไขดังที่ระบุไว้ด้านหลังทุกข้อ (ดูด้านหลังของเอกสารรับรองโครงการวิจัย)

พหูนํ ปณฺฑิโต ชีเว

ประวัติผู้เขียน

ชื่อ	นายธงไชย พ้องเสียง
วันเกิด	23 มิถุนายน 2533
สถานที่เกิด	โรงพยาบาลเสลภูมิ
สถานที่อยู่ปัจจุบัน	11 หมู่ที่ 3 ตำบลทุ่งเขาหลวง อำเภอทุ่งเขาหลวง จังหวัดร้อยเอ็ด 45170
สถานที่ทำงานปัจจุบัน	โรงพยาบาลทุ่งเขาหลวง จังหวัดร้อยเอ็ด
ประวัติการศึกษา	พ.ศ.2556 ปริญญาวิทยาศาสตรบัณฑิต (วท.บ.) สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏมหาสารคาม พ.ศ.2565 ปริญญาวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม

พหูนํ ปณฺฑิโต ชีเว