การเปรียบเทียบประสิทธิภาพการจำแนกผู้ป่วยโรคเบาหวานโดยใช้เทคนิคการแปลงข้อมูลสำหรับเทคนิคการทำเหมืองข้อมูล

อฑิตยา กะการดี[1], ไก้รุ่ง เฮงพระพรหม*[2] และ เกล้ากัลยา ศิลาจันทร์[3]

[1,2]สาขาวิชาวิทยาการข้อมูล คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม
[3]สาขาเทคโนโลยีคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม
* kairung2011.heng@gmail.com

บทคัดย่อ

วัตถุประสงค์ของงานวิจัยนี้คือการสร้างแบบจำลองการจำแนกประเภทผู้ป่วยโรคเบาหวานจากชุดข้อมูลที่ผ่านการแปลงข้อมูลในรูปแบบ Min-Max, Mean, Z-score, Root เพื่อทำการเปรียบเทียบว่าข้อมูลที่แปลงมาแล้วนั้นข้อมูลไหนเหมาะสมกับเทคนิคการจำแนกข้อมูลที่ให้ค่าความแม่นยำในการจำแนกที่ดีที่สุด ด้วยการเปรียบเทียบประสิทธิภาพของแบบจำลองของเทคนิคเหมืองข้อมูล 4 ประเภท ได้แก่ ด้วยเทคนิคคือ Neural Network, Decision tree และ k – nearest neighbor. จากการทดลองพบว่านิวรอลเน็ตเวิร์มีประสิทธิภาพการจำแนกข้อมูลแม่นยำสูงสุดคิดเป็นร้อยละ 75.13%

คำสำคัญ: การแปลงข้อมูล การจำแนกประเภท นิวรอลเน็ตเวิร์ก ต้นไม้ตัดสินใจ เพื่อนบ้านใกล้สุด k ตัว

A Comparison efficiency of classification of diabetic patients using data transformation techniques for data mining techniques

Athitaya Kakandee[1], Kairung Hengpraprohm[2] and Klaokanlaya Silachan[3]

[1,2]Data Science Department, Faculty of Science and Technology, Nakhon Pathom Rajabhat University

[3]Technologycomputer Department, Faculty of Science and Technology, Nakhon Pathom Rajabhat University

* kairung2011.heng@gmail.com

Abstract

The objective of this study was to built a classification model for diabetes patients from the transformed datasets using Min-Max, Mean, Z-score and Root formats to compare whether the transformed data were diabetic. Which is suitable for the classification technique that provides the best classification accuracy? By comparing the model efficiency of 4 types of data mining techniques, namely, Neural Network, Decision tree and k – nearest neighbor. From the experiment, it was found that the neural network had the highest efficiency in data classification accuracy is 75.13%.

Keywords: Transformation, Classification, Neural Network, Decision Tree, K-nearest neighbor

1. บทนำ

โรคเบาหวานเป็นสิ่งที่ส่งผลกระทบต่อปัญหาสุขภาพอื่นๆ มากมายทั้ง ฟัน เหงือก ตา ไต หัวใจ และหลอดเลือดแดง สาเหตุของโรคเบาหวานมีหลายปัจจัยร่วมกันทั้งปัจจัยทางพันธุกรรม (genetic factor) และปัจจัยสิ่งแวดล้อม (environmental factor) เป็นกลุ่มโรคเกี่ยวกับการเผาผลาญอาหารซึ่งเป็นภาวะที่ร่างกายมีระดับน้ำตาลในเลือดสูงเป็นเวลานาน เนื่องจากขาดฮอร์โมนอินซูลิน หรือประสิทธิภาพของอินซูลินลดลง ซึ่งมีผลต่อการทำงานของระบบต่างๆในร่างกาย รวมถึงอินซูลินเป็นฮอร์โมนสำคัญตัวหนึ่งของร่างกาย เพราะร่างกายคนเรานั้นรับประมานอาหารเข้าไปทุกวัน มีการเปลี่ยนแป้ง , โปรตีนให้เป็นน้ำตาล หากไม่มีอินซูลินก็จะส่งผลให้ร่างกายไม่สามารถนำน้ำตาลไปใช้เป็นพลังงานให้ส่วนต่างๆ ของร่างกายได้ และยังทำให้มีระดับน้ำตาลในเลือดสูงเช่นกัน อีกทั้งยังทำให้เกิดโรคเบาหวานได้ อินซูลินสร้างและหลั่งจากเบต้าเซลล์ของตับอ่อน ทำหน้าที่เป็นตัวพาน้ำตาลกลูโคสเข้าสู่เนื้อเยื่อต่างๆ ของร่างกาย เพื่อเผาผลาญเป็นพลังงานในการดำเนินชีวิต เนื่องจากผู้ที่เป็นโรคเบาหวานบางรายไม่ทราบว่าเป็นโรคเบาหวาน จึงไม่ได้ดูแลตนเองให้ระดับน้ำตาลในเลือดอยู่ในเกณฑ์ ซึ่งการที่ระดับน้ำตาลในเลือดสูงเป็นระยะเวลานาน จะส่งผลให้เกิดภาวะแทรกซ้อนต่อเท้า ไต ตา ระบบประสาท หัวใจและหลอดเลือด สมอง จึงส่งผลทำให้เกิดอาการป่วยและตายก่อนวัยอันสมควร [1][2]

โรคเบาหวาน คือ โรคเรื้อรังที่เป็นปัญหาสำคัญทางด้านสาธารณสุขของโลก เป็นภัยคุกคามที่ลุกลามอย่างรวดเร็วไปทั่วโลก ส่งผลกระทบต่อการพัฒนาทางเศรษฐกิจอย่างมาก จากข้อมูลสมาพันธ์เบาหวานนานาชาติ(international diabetes federation : IDF) ได้รายงานว่าในปัจจุบันทั่วโลกมีผู้เสียชีวิตด้วยโรคเบาหวาน ๔ ล้านคนต่อปี เฉลี่ย ๘ วินาทีต่อ ๑ คน สำหรับผู้เป็นเบาหวานพบมากกว่า ๓๐๐ ล้านคนและพบว่าคนที่อยู่ในประเทศที่มีรายได้ต่ำและปานกลางมีโอกาสเป็นเบาหวานเร็วกว่าคนที่อยู่ในประเทศที่มีรายได้สูง ๑๐ - ๒๐ ปี โดยพบมากขึ้นในวัยทำงาน เป็นความผิดปกติที่ร่างกายไม่สามารถนำน้ำตาลในเลือดไปใช้เป็นพลังงานได้ตามปกติ ทำให้มีระดับน้ำตาลในเลือดสูง ในปี พ.ศ. 2559 มีผู้ที่เป็นเบาหวาน 422 ล้านคนจากทั่วโลก นับเป็นอัตราส่วน 1 ใน 11 คน โดยพบว่าผู้ป่วย 1 ใน 2 คน ยังไม่ได้รับการวินิจฉัยว่าเป็นเบาหวาน และในปี พ.ศ.2564 กรมควบคุมโรค กระทรวงสาธารณสุข พบว่าโรคเบาหวานยังเป็นปัญหาในระดับโลกมาอย่างต่อเนื่อง และเป็นหนึ่งในกลุ่มโรคไม่ติดต่อเรื้อรัง (NCDs) สถานการณ์โรคเบาหวานทั่วโลกมีผู้ป่วยจำนวน 463 ล้านคน และคาดการณ์ว่าในปี 2588 จะมีผู้ป่วยเบาหวานจำนวน 629 ล้านคน สำหรับประเทศไทยพบอุบัติการณ์โรคเบาหวานมีแนวโน้มเพิ่มขึ้นอย่างต่อเนื่อง มีผู้ป่วยรายใหม่เพิ่มขึ้นประมาณ 3 แสนคนต่อปี และมีผู้ป่วยโรคเบาหวานอยู่ในระบบทะเบียน 3.2 ล้านคน ของกระทรวงสาธารณสุข ก่อให้เกิดการสูญเสียค่าใช้จ่ายในการรักษาด้านสาธารณสุขอย่างมหาศาล เฉพาะเบาหวานเพียงโรคเดียวทำให้สูญเสียค่าใช้จ่ายในการรักษาเฉลี่ยสูงถึง 47,596 ล้านบาทต่อปี โรคเบาหวานยังคงเป็นสาเหตุหลักที่ก่อให้เกิดโรคอื่นๆ ในกลุ่มโรค NCDs อีกมากมาย อาทิ โรคหัวใจ โรคหลอดเลือดสมอง โรคความดันโลหิตสูง และโรคไต ฯลฯ [1] [2]

จากความสำคัญของโรคดังกล่าว ผู้วิจัยจึงมีแนวคิดที่จะศึกษาเกี่ยวกับการจำแนกข้อมูลโรคเบาหวานโดยผ่านกระบวนการแปลงข้อมูลเป็นข้อมูลปกติ (data transformation) เนื่องการที่จะได้มาด้วยข้อมูลต่างๆนั้น ข้อมูลที่รวบรวมมาอาจมีความไม่เป็นระเบียบ มักเกิดปัญหาข้อมูลแต่ละตัวแปรมีค่าแตกต่างกันหากนำข้อมูลที่ไม่เป็นระเบียบนำมาทำการวิเคราะห์ข้อมูล อาจจะมีผลให้ผลลัพธ์ของการวิเคราะห์ข้อมูลเกิดการคลาดเคลื่อนได้ และศึกษาการสร้างแบบจำลองการจำแนกผู้ป่วยโรคเบาหวานจากชุดข้อมูลที่ผ่านการทรานฟอร์มโดยใช้เทคนิคเหมืองข้อมูลที่มีประสิทธิภาพในการจำแนกด้วยเทคนิควิธีต่างๆ เพื่อให้เกิดการเรียนรู้ที่ดีที่สุดสำหรับการจำแนกประเภทผู้ป่วยโรคเบาหวาน และเปรียบเทียบประสิทธิภาพของเทคนิควิธีที่หมาะสมกับข้อมูลโรคเบาหวาน [3]

ซึ่งผู้วิจัยได้ศึกษาข้อมูลงานวิจัยที่เกี่ยวกับการเปรียบเทียบประสิทธิภาพของโรคเบาหวานด้วยเทคนิคการทำเหมืองข้อมูลจากงานวิจัยหลายๆ แหล่ง พบว่ามีการใช้วิธีการจำแนกประเภทข้อมูลด้วยเทคนิคเหมืองข้อมูลในโดเมนที่แตกต่างกัน และมีวิธีการทำการคิด และการวิเคราะห์ที่แตกต่างกัน จึงทำการวิจัยจากแบบจำลองการจำแนกประเภทผู้ป่วยโรคเบาหวานจากชุดข้อมูลที่ผ่าน

การแปลงข้อมูลในรูปแบบ Min-Max, Mean, Z-score, Root เพื่อทำการเปรียบเทียบว่าข้อมูลที่แปลงมาแล้วนั้นข้อมูลไหนเหมาะสมกับเทคนิคการจำแนกข้อมูลที่ให้ค่าความแม่นยำในการจำแนกที่ดีที่สุด ด้วยการเปรียบเทียบประสิทธิภาพของแบบจำลองของเทคนิคเหมืองข้อมูล 4 ประเภท ได้แก่ ด้วยเทคนิคคือ Neural Network, Decision tree และ k – nearest neighbor เพื่อเพิ่มประสิทธิภาพในการวิเคราะห์โรคเบาหวานให้มีความแม่นยำขึ้น

2. วัตถุประสงค์การวิจัย

2.1 เพื่อใช้เทคนิคเหมืองข้อมูลในการแปลงชุดข้อมูลผู้ป่วยโรคเบาหวานให้อยู่ในรูปแบบที่เหมาะสมกับเทคนิคที่นำมาใช้ใช้จำแนกประเภทข้อมูล

2.2 เพื่อเปรียบเทียบประสิทธิภาพของการจำแนกประเภทผู้ป่วยโรคเบาหวานจากชุดข้อมูลการแปลงข้อมูลในรูปแบบต่างๆ

3. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

3.1 ทฤษฎีที่เกี่ยวข้อง

3.1.1 เหมืองข้อมูล

เหมืองข้อมูล (data Mining) เป็นเทคนิคในการวิเคราะห์ข้อมูลอย่างหนึ่ง ซึ่งมาจากคำว่า เหมืองข้อมูล นั่นคือ เป็นการค้นหาสิ่งที่มีประโยชน์จากฐานข้อมูลที่มีขนาดใหญ่ เช่น ข้อมูลการซื้อขายสินค้าในซุปเปอร์มาร์เก็ตต่าง ๆ โดยข้อมูลเหล่านี้จะเก็บจากรายการสินค้าที่ลูกค้าซื้อในแต่ละครั้ง โดยเมื่อทำการวิเคราะห์ข้อมูลด้วยเทคนิค Data Mining แล้วจะได้สิ่งที่เป็นประโยชน์ [3]

3.1.2 วิธีการเพื่อนบ้านใกล้ที่สุด

ขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor Algorithm : Knn) เป็นวิธีที่ใช้ในการจัดแบ่งคลาส โดยเทคนิคนี้จะตัดสินใจว่า คลาสใดที่จะแทนเงื่อนไขหรือกรณีใหม่ๆ ได้บ้าง โดยการตรวจสอบจำนวนบางจำนวน ในขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด ของกรณีหรือเงื่อนไขที่เหมือนกันหรือใกล้เคียงกันมากที่สุด โดยจะหาผลรวม (Count Up) ของจำนวนเงื่อนไขหรือกรณีต่างๆ สำหรับแต่ละคลาส และกำหนดเงื่อนไขใหม่ๆ ให้คลาสที่เหมือนกันกับคลาสที่ใกล้เคียงกันมากที่สุด [4]

3.1.3 ต้นไม้การตัดสินใจ

ต้นไม้การตัดสินใจ (decision tree) เป็นเครื่องมือที่ช่วยให้วิเคราะห์เหตุการณ์ หรือสถานการณ์เพื่อการตัดสินใจได้อย่างเป็นระบบและรวดเร็ว ต้นไม้การตัดสินใจมีลักษณะเป็นกราฟรูปต้นไม้ ซึ่งแสดงที่ตั้งต้นที่มีรากและแขนงต่างๆแตกออกมาจากต้นไม้ไปในทิศทางเดียว จนกระทั่งนำไปสู่ข้อสรุปสำหรับการตัดสินใจได้ ต้นไม้การตัดสินใจมีประโยชน์ในการสรุปการตัดสินใจที่มีความซับซ้อนให้ง่ายต่อความเข้าใจ ปัจจุบันต้นไม้การตัดสินใจเป็นที่นิยมใช้ในงานหลายอย่าง เช่น การแพทย์ ธุรกิจ การเขียนโปรแกรม การสร้างเครื่องที่เรียนรู้ได้เอง การสร้างระบบผู้เชี่ยวชาญ ฯลฯ [4]

3.1.4 โครงข่ายประสาทเทียม

โครงข่ายประสาทเทียม (Artificial Neural Networks) คือ การสร้างคอมพิวเตอร์ที่มีแบบจำลองที่จำลองวิธีการทำงานของสมองมนุษย์ หรือเป็นการทำให้คอมพิวเตอร์รู้จักการคิดและการจดจำ หรือจะใช้โครงข่ายประสาทเทียมในการทำให้คอมพิวเตอร์รู้จักและเข้าใจภาษามนุษย์ หรือเรียกอีกอย่าว่า AI [4]

3.1.5 การแปลงรูปแบบข้อมูล

แปลงรูปแบบของข้อมูล หรือ การทำให้เป็นปกติ (Data Transformation) เป็นส่วนหนึ่งของขั้นตอนการเตรียมข้อมูล(Data preparation) ซึ่งเป็นการปรับเปลี่ยนรูปแบบของข้อมูลให้อยู่ในรูปแบบที่พร้อมนำไปใช้ในการประมวลผลหรือการวิเคราะห์ต่อไป [4] มีหลายเทคนิควิธีเช่น

3.1.5.1 Min-Max Normalization คือ การทำให้เป็นมาตรฐานต่ำสุด-สูงสุดเป็นสมการเชิงเส้นบนเพื่อทำให้เป็นชุดข้อมูลปกติ การทำให้เป็นมาตรฐานต่ำสุด-สูงสุดจะแปลงค่า d ของ P ถึง d' ที่อยู่ในช่วง [new_min(p),new_max(p)] โดยสูตรต่อไปนี้ [5]

$$d' = \frac{d - \min(p)}{\max(p) - \min(p)} \quad (1)$$

การทำให้เป็นมาตรฐานต่ำสุดสูงสุดจะรักษาความสัมพันธ์ระหว่างค่าข้อมูลดั้งเดิม.

3.1.5.2 Mean Normalization คือ Mean Normalization คล้ายกับ Rescaling ด้านบน แตกต่างกันที่ใช้ Mean แทน Min ทำให้ช่วงของ Output [-0.5, 0.5] มีทั้งบวกและลบ Balance กัน ตรงเลข 0 (ขยับ Mean มาตรง 0) [6]

$$x' = \frac{x - \text{average}(x)}{\max(x) - \min(x)} \quad (2)$$

3.1.5.3 Z-score Normalization

การทำให้เป็นมาตรฐานของคะแนน Z เรียกอีกอย่างว่าการทำให้เป็นมาตรฐานเป็นศูนย์ ในการทำให้เป็นมาตรฐานของคะแนน Z ค่าสำหรับแอตทริบิวต์ P จะถูกทำให้เป็นมาตรฐานโดยอิงตามค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของ P ค่า d ของ P จะถูกทำให้เป็นมาตรฐานเป็น d′ โดย t การทำให้เป็นมาตรฐานต่ำสุด-สูงสุดจะทำการแปลงเชิงเส้นในข้อมูลดั้งเดิม การทำให้เป็นมาตรฐานต่ำสุด-สูงสุดแมปค่า d ของ P ถึง d' ในช่วง [new_min(p),new_max(p)] การทำให้เป็นมาตรฐานต่ำสุด-สูงสุดคำนวณโดยสูตรต่อไปนี้: [5]

$$d' = \frac{d - mean(P)}{std(p)} \quad (3)$$

โดยที่ mean (p) = ค่าเฉลี่ยของแอตทริบิวต์ P และ std (p) = มาตรฐานการเบี่ยงเบนของแอตทริบิวต์ P

3.1.5.4 การแปลงโดยใช้รากที่สอง (Square Root Normalization)

การแปลงข้อมูลโดยใช้ [7]

$$y^*_{ij} = \sqrt{y_{ij}} \quad (4)$$

เมื่อ y_{ij} คือ ค่าสังเกตเดิม, y^*_{ij} คือข้อมูลที่ถูกแปลง

3.2 งานวิจัยที่เกี่ยวข้อง

[8] งานวิจัยนี้ทำเพื่อศึกษาการพยากรณ์ความเสี่ยงการเกิดโรคเบาหวาน โดยนำค่าปัจจัยเสี่ยงของผู้ป่วยโรคเบาหวานมาใช้ในการพยากรณ์ ซึ่งศึกษาและวิเคราะห์ปัจจัยเสี่ยงของโรคเบาหวาน โดยใช้ปัจจัยเฉพาะด้านจากข้อมูลผู้ป่วยโรงพยาบาลมหาสารคาม ตามทะเบียนผู้ป่วย จำนวน 5,000 คน ตั้งแต่ สิงหาคม พ.ศ. 2558 – ธันวาคม พ.ศ. 2559 เพื่อสร้างตัวแบบที่เหมาะสมที่สุดในการวิเคราะห์ปัจจัยเสี่ยงของโรคเบาหวาน เทคนิคที่นำมาใช้ในการพยากรณ์ความเสี่ยงของโรคเบาหวาน คือ วิธี Decision Tree ID3 ซึ่งข้อมูลที่วิจัยฉบับนี้เลือมา คือ การเก็บสภาพอากาศย้อนหลัง 14 วันเพื่อดูว่าในแต่ละวันจะมีการ จัดการแข่งขันเบสบอลขึ้นหรือไม่ ซึ่งจะมีแอตทริบิวต์ Outlook , Temperature , Humidity และ Windy จะเป็นแอตทริบิวต์ประเภททั่วไปที่ใช้ในการพิจารณาว่าถ้าค่าในแอตทริบิวต์เหล่านี้เป็นลักษณะไหนแล้วจึงมีการจัดแข่งเบสบอล ส่วนแอตทริบิวต์ที่เป็นคลาสคำตอบที่เราสนใจคือ แอตทริบิวต์ Play ผลการดำเนินการวิจัยครั้งนี้ ได้ตัวแบบในการพยากรณ์ด้วยอัลกอริทึม Decision Tree ID3 ซึ่งผลการประเมินประสิทธิภาพตัวแบบ จากการแบ่งข้อมูลทดสอบออกเป็น 5 ชุด ค่าความถูกต้อง (Accuracy) ได้ 69.45%

[9] วิจัยนี้ทำเพื่อการพัฒนาแบบจำลองปัจจัยที่มีผลต่อการเป็นโรคเบาหวานด้วยเทคนิคเหมืองข้อมูลแบบต้นไม้ตัดสินใจ โดยใช้ข้อมูลผู้เข้ารับบริการที่โรงพยาบาลด่านขุนทด จังหวัดนครราชสีมา ระหว่างปี 2550 – 2555 จำนวนทั้งสิ้น 4,402 ราย แบ่งข้อมูลสำหรับฝึกและทดสอบแบบจำลองออกเป็นร้อยละ 90:10 ตามลำดับ พัฒนาแบบจำลองด้วยอัลกอริทึม j48 ซึ่งเป็นเทคนิคแบบต้นไม้ตัดสินใจ ประเมินประสิทธิภาพของแบบจำลองด้วยค่าความแม่นตรง โดยอาศัยโปรแกรมประยุกต์ด้านเหมืองข้อมูล เพื่อค้นหาปัจจัยที่อาจเป็นสาเหตุของการเกิดโรคเบาหวานได้ ผู้วิจัยได้ดำเนินการพัฒนาแบบจำลองของปัจจัยที่มีผลต่อการเป็นโรคเบาหวานด้วยเทคนิคเหมืองข้อมูลแบบต้นไม้ตัดสินใจผลการพัฒนาแบบจำลอง จากผลการพัฒนาแบบจำลอง เมื่อเปรียบเทียบวิธีการทดสอบทั้ง 5 วิธีพบว่าวิธีที่ให้ค่าความแม่นตรงสูงสุดคือ วิธีทดสอบแบบแยกชุด 90:10 ที่ให้ค่าเท่ากับ 76.14% ดังนั้นเมื่อนำแบบจำลองที่ให้ค่าความแม่นตรงสูงสุด ใช้ทำนายข้อมูลชุดใหม่ที่เตรียมไว้400 รายการ ผลการทำนายพบว่าแบบจำลองที่ได้ทำนายถูกต้อง 315 รายการ คิดเป็น 78.75% และทายผิด 85 รายการ คิดเป็น 21.25%

[10] วิจัยนี้ทำเพื่อการสร้างแบบจำลองและการเปรียบเทียบประสิทธิภาพการจำแนกประเภทผู้ป่วยโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูลและการเลือกคุณลักษณะจากความสัมพันธ์ของข้อมูลและทำการเปรียบเทียบประสิทธิภาพของแบบจำลองของเทคนิคเหมืองข้อมูล 4 ประเภท ได้แก่ เนอีฟเบย์ (Naïve Bayes), ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine), เคเนียเรสเนเบอร์ (K-Nearest Neighbor), ต้นไม้ตัดสินใจ (Decision Tree) โดยทดสอบกับข้อมูลทางการแพทย์ผู้ป่วยโรคเบาหวาน จากข้อมูลชุดทดสอบของผู้ป่วยโรคเบาหวานจำนวน 768 คน ผลการทดลองสำหรับการทำนายลักษณะจำแนกของผู้ป่วยโรคเบาหวานโดยใช้แบบจำลองเหมืองข้อมูลที่สร้างโดยใช้เทคนิคซัพพอร์ทเวกเตอร์แมชชีน มีประสิทธิภาพการทำนายสูงสุดคิดเป็น 76.95% ซึ่งสามารถนำผลที่ได้ไปประยุกต์ใช้ในการสร้างระบบสนับสนุนการตัดสินใจในส่วนของคัดกรองโรคและแนวทางการรักษาของแพทย์และผู้ป่วย

[11] วิจัยนี้ทำเพื่อศึกษาเทคนิคการทำเหมืองข้อมูลที่เหมาะสมในการพยากรณ์ ความเสี่ยงของการเกิดโรคเบาหวาน โดยทำการศึกษาเทคนิคต่างๆ จากงานวิจัยที่มีการนำเทคนิคเหมืองข้อมูลมาใช้ในการพยากรณ์หาสาเหตุและความเสี่ยงในการเกิดโรค คือ เทคนิคต้นไม้ตัดสินใจ(Decision Tree) และโครงข่ายประสาทเทียมแบบแพร่กลับ(Backpropagation Neural Network: BPNN) ผลการทดลองวิจัยนี้เป็นการศึกษาเทคนิคการทำเหมืองข้อมูลเพื่อพยากรณ์ความเสี่ยงการเกิดโรคเบาหวานโดยผู้วิจัยได้ทำการศึกษางานวิจัยที่เกี่ยวข้อง เพื่อค้นหาเทคนิคที่เหมาะสมในการพยากรณ์และจากการศึกษาพบว่า เทคนิคต้นไม้ตัดสินใจ (Decision Tree) และโครงข่ายประสาทเทียมแบบแพร่กลับ(Backpropagation Neural Network: BPNN) มีค่าความถูกต้องและมีความเหมาะสมในการนำไปใช้สร้างแบบจำลองพยากรณ์ความเสี่ยงของการเกิดโรคเบาหวาน

[12] วิจัยนี้ทำเพื่อศึกษาหาแนวทางการหาปัจจัยที่เป็นเหตุนำไปสู่การเกิดโรคเบาหวานชนิดที่2 และเพื่อศึกษาความสัมพันธ์ของข้อมูล เพื่อนำไปพัฒนาเป็นตัวแบบพยากรณ์โรคเบาหวานชนิดที่ 2 จากการศึกษากลุ่มตัวอย่างผู้ป่วยที่เป็นโรคเบาหวานชนิดที่2 ที่เข้ารับการรักษาที่โรงพยาบาลรัฐบาลแห่งหนึ่ง ในเขตกรุงเทพมหานคร ระหว่างเดือนมกราคม 2552 - ตุลาคม 2557 จำนวน 46,000 รายงานวิจัยนี้ใช้เทคนิคการทำเหมืองข้อมูล (Data Mining) ตามกรอบ CRISP-DM โดยผู้วิจัยได้เริ่มการวิเคราะห์จากการสร้างแบบจำลองการแบ่งกลุ่ม (Clustering) ด้วยวิธีSimple K-Means เพื่อใช้ในการจัดกลุ่มผู้ป่วยโรคเบาหวานชนิดที่2 จากนั้นนำผลที่ได้มาวิเคราะห์การถดถอยโลจิสติกทวิ (Binary Logistic Regression Analysis) เพื่อพยากรณ์โอกาสที่จะนำไปสู่พฤติกรรมของผู้ป่วยโรคเบาหวานชนิดที่2 การวิจัยในครั้งนี้ก่อให้เกิดประโยชน์ในหลายๆ ด้าน เช่น ด้านสถาบันการแพทย์จะทำให้ผู้เชี่ยวชาญทางการแพทย์ พยาบาล รวมทั้งเจ้าหน้าที่ที่เกี่ยวข้องสามารถนำความรู้ที่ได้ จากการวิเคราะห์ไปใช้ในการวินิจฉัยโรค รวมทั้งเพื่อหาแนวทางในการรักษาโรคต่อไป ด้านผู้ป่วยจะช่วยผู้ป่วยทราบถึงโอกาสที่จะเป็น โรคเบาหวานชนิดที่2 เพื่อจะได้หาทางป้องกันได้อย่างทันท่วงที

4. วิธีดำเนินการวิจัย

4.1 กรอบแนวคิดในการวิจัย

การดำเนินการวิจัย ประกอบด้วยกรอบแนวคิดในการวิจัยดังนี้

4.1.1 ทำการศึกษาลักษณะชุดข้อมูล diabetes

4.1.2 ทำข้อมูลให้สมบูรณ์ (Data Cleaning) ตรวจสอบแก้ไขข้อมูลที่ไม่เกี่ยวข้องออกไป

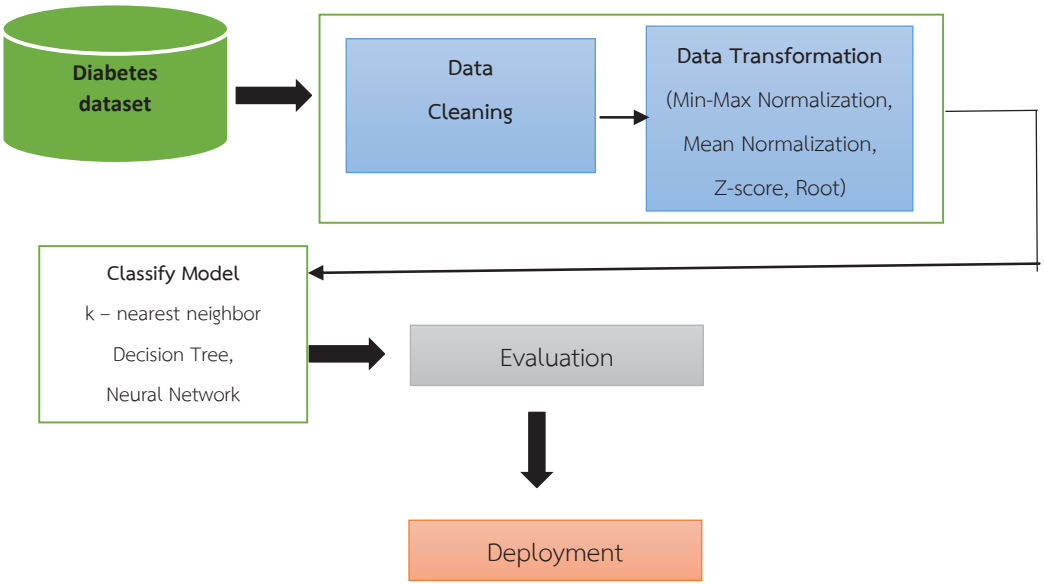
4.1.3 แปลงข้อมูลให้เหมาะสมสำหรับการใช้งาน (Transformation) การแปลงข้อมูลให้เหมาะสมสำหรับการวิเคราะห์งาน ในการวิจัยนี้จะเลือกการแปลงข้อมูลด้วย Rescaling หรือ Min-Max Normalization และ Mean Normalization , Z-score , Root เพื่อทำการเปรียบเทียบว่าข้อมูลที่แปลงมาแล้วนั้นข้อมูลไหน เหมาะสมกับเทคนิคเหมืองข้อมูลที่เลือกมาข้างต้น

4.1.4 จำแนกประเภทกลุ่มโรค(Classify) จากนั้นนำข้อมูลที่ผ่านการแปลงข้อมูลนำเข้าสู่วิธีการจำแนกข้อมูลด้วยการใช้ Data Mining Techniques คือ k – nearest neighbor, Decision Tree และ Neural Network เพื่อเปรียบเทียบประสิทธิภาพการจำแนกที่มีค่าความแม่นยำ ด้วยการใช้ซอฟต์แวร์ Weka ในการจำแนกดังกล่าว

4.1.5 ประเมินผล(Evaluation) เพื่อนำผลที่ได้จากการนำเข้าWeka มาพิจารณาว่าเทคนิคเหมืองข้อมูลชนิดไหนมีประสิทธิภาพที่เหมาะสมที่สุด

4.1.6 นำแบบจำลองใปใช้งาน (Deployment) นำผลการวิเคราะห์ของแบบจำลองที่ทำการเปรียบเทียบประสิทธิภาพเพื่อได้แบบจำลองเทคนิคที่มีความน่าเชื่อถือ และสามารถนำผลที่ได้มาใช้ให้เกิดประโยชน์เกี่ยวกับทางการแพทย์ได้ต่อไป

สามารถนำเสนอในรูปแบบแผนภาพแบบจำลองกรอบแนวคิดการวิจัย ดังภาพที่ 1 ดังนี้



ภาพที่ 1 กรอบแนวคิดในการวิจัย

4.2 ชุดข้อมูลสำหรับการวิจัย

ชุดข้อมูล diabetes ที่นำมาใช้ในงานวิจัยได้นำมาจากแหล่งข้อมูลจากเวบไซต์ของ Kaggle.com ในชุดของ pima_diabetes เป็นชุดข้อมูลเกี่ยวกับโรคเบาหวาน ภายในจะมี 9 คอลัมน์ และ768 แถว โดยข้อมูลมีคุณลักษณะประกอบด้วย รายละเอียดคือ การตั้งครรภ์ (preg หรือ pregnant), พลาสมา(plas หรือ plasma) , ความดัน(pres หรือ Pressure), ผิว (Skin), อินซูลิน(insu หรือ insulin), ดัชนีมวลร่างกาย(mas) , ค่าเบาหวานในเด็ก(pedi หรือ Pediatrics) , อายุ(age) และ ค่าในการทดสอบ(tested_ positive, tested _negative) โดยมีตัวอย่างโครงสร้างและข้อมูล ดังตารางที่ 1

ตารางที่ 1 แสดงโครงสร้างและข้อมูลโรคเบาหวาน(pima diabetes)

pregnant	Plass	pres	skin	insu	mass	pedi	age	class
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1

4.3 การประเมินผลการวิจัย

การวิจัยนี้ใช้วิธีการประเมินประสิทธิภาพด้วยค่าความแม่นยำในการจำแนก (Accuracy) คือ มาตรวัดค่าความแม่นตรง คือ ค่าที่บอกถึงความแม่นยำในการจำแนกข้อมูล จากสมการ [13]

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

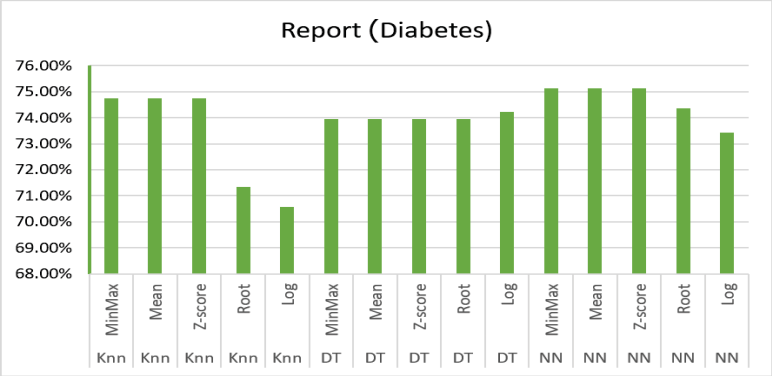
โดยที่ TP คือค่า True Positive, TN คือค่า True Negative,
FP คือค่า False Positive, FN คือค่า False Negative

5. ผลการวิจัย

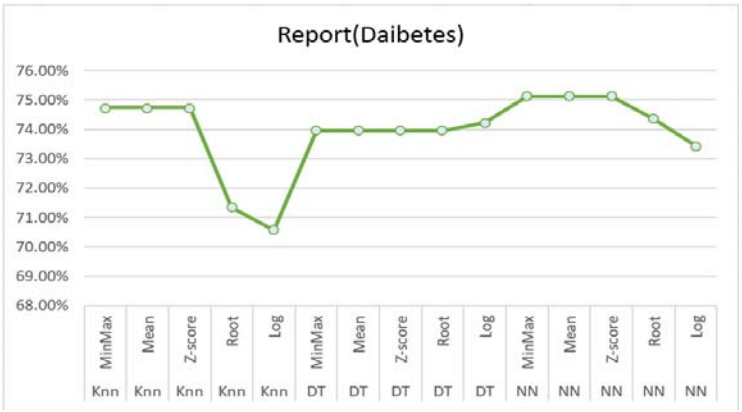
ผลการแปลงชุดข้อมูลและ จำแนกประเภทข้อมูลจากชุดข้อมูลการแปลงแต่ละเทคนิค ผ่านเทคนิคการจำแนกขอมูล โดยแบงขอมูลในการเรียนรู (Training Data) และขอมูลในการทดสอบ (Testing Data) ตามวิธีการ k-fold cross validation โดยกำหนด k=10 [1] แสดงผลในรูปแบบตารางที่ 2 และภาพที่ 2-3

ตารางที่ 2 เปรียบเทียบแสดงค่าความแม่นยำในการจำแนกประเภทจากเทคนิค กับชุดข้อมูลการแปลงข้อมูล

เทคนิค	tranfrom	ค่า accuracy
Knn	MinMax	74.74%
Knn	Mean	74.74%
Knn	Z-score	74.74%
Knn	Root	71.35%
Knn	Log	70.57%
DT	MinMax	73.96%
DT	Mean	73.96%
DT	Z-score	73.96%
DT	Root	73.96%
DT	Log	74.22%
NN	MinMax	75.13%
NN	Mean	75.13%
NN	Z-score	75.13%
NN	Root	74.35%
NN	Log	73.44%



ภาพที่ 2 ผลเปรียบเทียบแสดงค่าความแม่นยำในการจำแนกประเภทจากเทคนิค กับชุดข้อมูลการแปลงข้อมูล ในรูปแบบกราฟแท่ง



ภาพที่ 3 ผลการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลจากชุดข้อมูลการทรานฟอร์ม จะแสดงในรูปแบบกราฟเส้น

จากผลการเปรียบเทียบประสิทธิภาพของชุดข้อมูล Diabetes ด้วยเทคนิค Neural Network ผ่านการแปลงข้อมูลด้วยวิธี Rescaling หรือ Min-Max Normalization เท่ากับ 75.13% ,Mean Normalization เท่ากับ 75.13% เทคนิคDecision tree ผ่านการแปลงข้อมูลด้วยวิธี Rescaling หรือ Min-Max Normalization เท่ากับ 73.96% , Mean Normalization เท่ากับ 73.96% เทคนิคk – nearest neighbor ผ่านการแปลงข้อมูลด้วยวิธีRescaling หรือ Min-Max Normalization เท่ากับ 74.74% และ Mean Normalization เท่ากับ 74.74% ดังนั้นจึงสรุปผลได้ว่า เทคนิค Neural Network ผ่านการแปลงข้อมูลด้วยวิธีRescaling หรือ Min-Max Normalization และMean Normalization เท่ากับ 75.13% มีประสิทธิภาพดีที่สุด

6. สรุปผล

ผลการวิจัยนี้ได้ทำการทรานฟอร์มข้อมูลด้วยวิธีการ Min-Max, Mean, Z-core เพื่อทำชุดข้อมูลให้เป็นปกติและทำการเปรียบเทียบการจำแนกกลุ่มจากชุดข้อมูลที่แปลงให้เป็นปกติทั้ง 4 วิธีการด้วยเทคนิคการจำแนกกลุ่มจำนวน 4 เทคนิควิธี คือ Neural Network, Decision tree และ k – nearest neighbor จากชุดข้อมูลการทรานฟอร์ม ซึ่งผลการประเมินประสิทธิภาพตัวแบบ คือ Neural Network และชุดข้อมูลการทรานฟอร์มด้วย Min-Max Normalization และ,Mean Normalization ให้ค่าความแม่นยำมากที่สุดเท่าๆ กัน นั่นคือ ค่าความถูกต้อง (Accuracy) ได้ 75.13% จึงสรุปได้ว่า Neural Network เป็นตัวแบบที่เหมาะสมที่สุดที่จะนำมาใช้จำแนกข้อมูลผู้ป่วยโรคเบาหวานชุดนี้

7. ข้อเสนอแนะ

ในการแปลงข้อมูลให้เป็นข้อมูลปกติเพื่อลดความคลาดเคลื่อนในการนำข้อมูลไปวิเคราะห์ หากชุดข้อมูลมีจำนวนมากขึ้นและต้องการเพิ่มความเร็วในการแปลงข้อมูลสามารถพัฒนาโปรแกรมเพื่อทำการอ่านชุดข้อมูลและแปลงข้อมูลให้อยู่ในรูปแบบตามเทคนิคของการแปลงข้อมูลต่างๆ และในส่วนการจำแนกประเภทข้อมูลสำหรับงานวิจัยนี้ใช้โปรแกรม Weka เพื่อนำมาประยุกต์ใช้ในการจำแนกสามารถทดลองด้วยเทคนิควิธีการอื่นเพื่อเปรียบเทียบกับ Neural Network จากชุดข้อมูลชุดนี้เพื่อจะได้เทคนิควิธีการจำแนกที่มีหาค่าความถูกต้องมากที่สุด

8. เอกสารอ้างอิง

[1] จิรพรรณ ผิวนวล และประทุม เนตรินทร.**พฤติกรรมการดูแลตนเองของผู้ป่วยเบาหวานที่ควบคุมระดับน้ำตาลในเลือดไม่ได้ โรงพยาบาลส่งเสริมสุขภาพต้าบลบางแก้วใน ต้าบลบางแก้ว อ้าเภอละอุ่น จังหวัดระนอง.** วารสารวิทยาลัยพยาบาลพระจอมเกล้า จังหวัดเพชรบุรี. ปีที่ 1 ฉบับที่ 2 , 2561.

[2] สมาคมโรคเบาหวานแห่งประเทศไทยฯ. **วารสารเบาหวาน.** ปีที่ 53, ฉบับที่ 1.

[3] ปิยวรรณ นิลถนอม, ธนพร มาลัย และ สายชล สินสมบูรณ์ทอง.การเปรียบเทียบประสิทธิภาพการทำนายผลการแปลงข้อมูลในการจำแนกด้วยเทคนิคการทาเหมืองข้อมูล.*Thai Journal of Science and Technology. Vol. 10 • No. 1 • 2021.*

[4] สัญฌา พันธุ์แพง.(2563). **การประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลสำหรับการพยากรณ์การศึกษาต่อนักศึกษาใหม่ระดับปริญญาตรี ในมหาวิทยาลัยราชภัฏเชียงใหม่.** แม่ฮ่องสอน : วิทยาลัยแม่ฮ่องสอน มหาวิทยาลัยราชภัฏเชียงใหม่.

[5] Yogendra K. J., & Santosh, K. B.. **(2013). Min Max Normalization Based Data Perturbation Method for Privacy Protection**. International Journal of Computer and Communication Technology. IIM : Bhubaneswar.

[6] Surapong Kanoktipsatharporn. (2562). **Normalization คืออะไร ปรับช่วงข้อมูล Feature Scaling ด้วยวิธี** Normalization, Standardization. ค้นเมื่29 พฤษภาคม 2565 จาก https://www.bualabs.com/ archives/2100 /what-is-normalization-feature-scaling-rescaling-normalization- standardization-feedforward-train-machine-learning-preprocessing-ep-2/.

[7] ทรงพล รดิศพงศ์.(2555)**. การแปลงข้อมูลผลการวิจัยโดยวิธีทางสถิติ.** วารสารกรมวิทยาศาสตร์บริการ ปีที่ 60 ฉบับที่ 189.

[8] รักถิ่น เหลาหา. (2560). **รายงานการวิจัยเรื่อง การพยากรณ์ความเสี่ยงการเกิดโรคเบาหวานโดยเทคนิคการทำเหมืองข้อมูล : กรณีโรงพยาบาลมหาสารคาม.** มหาสารคาม: มหาวิทยาลัยราชภัฏมหาสารคาม.

[9] วนิดา พงษ์สงวน, ทิพยา ถินสูงเนิน และมาโนช ถินสูงเนิน. (2561). **รายงานการวิจัยเรื่อง การพัฒนาแบบจำลองปัจจัยที่มีผลต่อการเป็นโรคเบาหวานด้วยเทคนิคต้นไม้ตัดสินใจ.** นครราชสีมา: มหาวิทยาลัยราชภัฏนครราชสีมา.

[10] รุ่งโรจน์ บุญมา และนิเวศ จิระวิชิตชัย. (2562). **รายงานการวิจัยเรื่อง การจำแนกประเภทผู้ป่วยโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูลและการเลือกคุณลักษณะจากความสัมพันธ์ของข้อมูล.** กรุงเทพมหานคร: มหาวิทยาลัยศรีปทุม.

[11] จตุพล จิตติยพล, ณัฐิสรา ชูลิขิต, พัชรินทร์ พลเยี่ยม และนลัทพร โอษฐิเวช. (2561). **รายงานการวิจัยเรื่อง การศึกษาเทคนิคการทำเหมืองข้อมูลเพื่อพยากรณ์ความเสี่ยงการเกิดโรคเบาหวาน.** ขอนแก่น: มหาวิทยาลัยขอนแก่น.

[12] ชนิดา เสือเปีย และกมล เกียรติเรืองกมลา. (2560). **รายงานการวิจัยเรื่อง การศึกษาสาเหตุการเกิดโรคเบาหวานชนิดที่ 2 ด้วยเทคนิคการทำเหมืองข้อมูลกรณีศึกษา โรงพยาบาลรัฐบาลแห่งหนึ่ง.** กรุงเทพมหานคร: วิทยาลัยนวัตกรรมมหาวิทยาลัยธรรมศาสตร์.

[13] พัฒนพงษ์ ดลรัตน์, จารี ทองคำ.(2560). **การเปรียบเทียบประสิทธิภาพของแบบจำลองในการพยากรณ์ความสำเร็จการศึกษาของนักเรียนระดับประกาศนียบัตรวิชาชีพ. วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยมหาสารคาม.** มหาสารคาม: มหาวิทยาลัยมหาสารคาม.