การจำแนกประเภทผู้ป่วยโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูล และการเลือกคุณลักษณะจากความสัมพันธ์ของข้อมูล

Classification of Diabetes Patient by Using Data Mining Techniques and Correlation Based Feature Selection

รุ่งโรจน์ บุญมา* และ นิเวศ จิระวิชิตชัย

วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ
มหาวิทยาลัยศรีปทุม
ถนน พหลโยธิน แขวง เสนานิคม เขตจตุจักร กรุงเทพมหานคร 10900

Submitted 7/7/2019 ; Revised 7/8/2019 ; Accepted 6/9/2019

บทคัดย่อ

วัตถุประสงค์ของงานวิจัยนี้คือการสร้างแบบจำลองการจำแนกประเภทผู้ป่วยโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูลและการเลือกคุณลักษณะจากความสัมพันธ์ของข้อมูลและทำการเปรียบเทียบประสิทธิภาพของแบบจำลองของเทคนิคเหมืองข้อมูล 4 ประเภท ได้แก่ เนอีฟเบย์, เคเนียเรสเนเบอร์, ต้นไม้ตัดสินใจ และซัพพอร์ทเวกเตอร์แมชชีน จากการทดลองพบว่าซัพพอร์ทเวกเตอร์แมชชีนมีประสิทธิภาพการทำนายสูงสุดคิดเป็น 76.95 สามารถนำผลที่ได้จากงานวิจัยนี้ไปประยุกต์ใช้ในการคัดกรองและสร้างระบบสนับสนุนการตัดสินใจในส่วนของแนวทางการรักษาของแพทย์ต่อไป

คำสำคัญ: เบาหวาน เหมืองข้อมูล ซัพพอร์ทเวกเตอร์แมชชีน

Abstract

The objective of this research is to create a classification model for diabetic patients by using Data Mining Techniques and correlation based feature selection and comparison of the efficiency of the models of 4 Data Mining Techniques, consisting of Naïve Bayes, Decision Tree, K-Nearest Neighbor and Support Vector Machine. The experimental results showed that support vector machine has the highest prediction efficiency of 76.95%. The results from the research can be applied to screening and establish a decision support system for the treatment of doctors.

Keyword: Diabetes, Data Mining, Support Vector Machine

***ผู้ประสานงาน (Corresponding Author)**
E-mail: rgb_@hotmail.co.th

1. บทนำ

ปัจจุบันในยุคของข้อมูลข่าวสาร องค์กรส่วนใหญ่มีข้อมูลที่ต้องจัดเก็บอยู่เป็นจำนวนมากมาย เช่น ระบบร้านค้าปลีก จะเก็บข้อมูลพนักงานในองค์กร ข้อมูลการซื้อขาย ข้อมูลสินค้า ข้อมูลลูกค้า เป็นต้น จะเห็นได้ว่า ยิ่งองค์กรหรือรูปแบบธุรกิจมีขนาดใหญ่เท่าไร ย่อมทำให้การเก็บสะสมข้อมูลสำหรับองค์กรต่าง ๆ มีจำนวนมากขึ้น การเก็บข้อมูลจำนวนมากเหล่านี้ลงในฐานข้อมูล เป็นวิธีที่นิยมใช้ในหลายองค์กร แต่ระบบการจัดการฐานข้อมูลทั่วไปไม่สามารถจัดการกับข้อมูลเหล่านี้ได้อย่างมีประสิทธิภาพ เนื่องจากใช้เวลานานในการดึงข้อมูลที่มีความสำคัญออกมาวิเคราะห์ ดังนั้นจึงได้เกิดเทคโนโลยีในการวิเคราะห์ข้อมูลที่มีความสำคัญออกมาจากแหล่งเก็บข้อมูลขนาดใหญ่ เรียกเทคโนโลยีนี้ว่า "การทำเหมืองข้อมูล" หรือ การขุดค้นข้อมูล (Data Mining)

หัวใจสำคัญของกระบวนการทำเหมืองข้อมูล คือส่วนของโปรแกรมที่ทำหน้าที่สังเคราะห์ความรู้ขึ้นมาจากข้อมูลจำนวนมากในฐานข้อมูล ส่วนสังเคราะห์ความรู้นี้เรียกว่า Learning algorithm ซึ่งมีผู้เสนอแนวคิดและพัฒนาอัลกอริทึมส่วนนี้ขึ้นเป็นจำนวนมาก ได้แก่ อัลกอริทึมที่ใช้หลักการของการสร้างต้นไม้ตัดสินใจ (Decision Tree) อัลกอริทึมที่ใช้หลักการทางสถิติและทฤษฎีของเบยส์ (Naïve Bayes) อัลกอริทึมที่ใช้หลักการของโครงข่ายประสาทเทียม (Neural Network) และอัลกอริทึมอื่น ๆ อีกมาก ปัจจุบันได้มีนักคอมพิวเตอร์ทดสอบเปรียบเทียบความสามารถของอัลกอริทึมแต่ละประเภท เพื่อค้นหาว่าอัลกอริทึมใดมีความสามารถสูงที่สุด ผลการทดสอบส่วนใหญ่ปรากฏว่าไม่มีอัลกอริทึมใดที่ทำงานได้ดีที่สุดในข้อมูลทุกประเภท ทั้งนี้เนื่องจากข้อมูลแต่ละประเภทมีลักษณะเฉพาะตัวที่ต่างกัน เช่น ข้อมูลทางการแพทย์จะต่างจากข้อมูลด้านกฎหมาย และต่างจากข้อมูลด้านธุรกิจ ดังนั้นจึงไม่มีอัลกอริทึมใดที่ดีที่สุดสำหรับข้อมูลทุกประเภท [1]

โรคเบาหวาน (Diabetes mellitus (DM) หรือทั่วไปว่า Diabetes) เป็นกลุ่มโรคเกี่ยวกับการเผาผลาญอาหารซึ่งมีระดับน้ำตาลในเลือดสูงเป็นเวลานาน น้ำตาลในเลือดสูงก่อให้เกิดอาการปัสสาวะบ่อย กระหายน้ำและความหิวเพิ่มขึ้น หากไม่ได้รับการรักษา เบาหวานอาจก่อให้เกิดอาการแทรกซ้อนจำนวนมากภาวะแทรกซ้อนเฉียบพลัน เบาหวานเป็นโรคเรื้อรังที่เป็นปัญหาสำคัญทางด้านสาธารณสุขของโลก เป็นภัยคุกคามที่ลุกลามอย่างรวดเร็วไปทั่วโลก ส่งผลกระทบต่อการพัฒนาทางเศรษฐกิจอย่างมาก จากปัญหาดังกล่าวผู้วิจัยจึงมีแนวคิดที่จะศึกษา ค้นหาลักษณะ และเปรียบเทียบประสิทธิภาพอัลกอริทึมสังเคราะห์ความรู้ ที่เหมาะสมกับข้อมูลด้านการแพทย์ เพื่อศึกษาการสร้างแบบจำลองการจำแนกผู้ป่วยโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูลที่มีประสิทธิภาพการสังเคราะห์ หรือการเรียนรู้ที่ดีที่สุดสำหรับการจำแนกประเภทผู้ป่วยโรคเบาหวาน และศึกษาวิธีการเปรียบเทียบประสิทธิภาพอัลกอริทึมนั้นด้วยเทคนิคต่าง ๆ เพื่อจะเพิ่มขีดความสามารถการวิเคราะห์โรคให้แม่นยำมากขึ้น โดยงานวิจัยครั้งนี้ได้ใช้ข้อมูลจาก UCI ซึ่งเป็นข้อมูลเปิดสำหรับนำไปทดสอบประสิทธิภาพอัลกอริทึมทางด้านเหมืองข้อมูล ได้แก่ฐานข้อมูล Diabetes Data จากกลุ่มตัวอย่างจำนวน 768 คน [6] ซึ่งจะช่วยส่งผลดีต่อการทำงานของแพทย์ คือ ช่วยลดระยะเวลาในการประเมินผลการเข้ามารักษา และผู้ป่วยสามารถทำได้ด้วยตัวเองโดยการประเมินจากผลน้ำตาลในเลือดและการใช้ยาทำให้เกิดความสะดวกสบายให้แก่ทั้งผู้ป่วยและแพทย์เอง ซึ่งสามารถนำไปประยุกต์ใช้ในการสร้างระบบสนับสนุนการตัดสินใจในส่วนของแนวทางการจำแนกของแพทย์และผู้ป่วยได้ต่อไป [8]

1.1 กรอบแนวคิดในการวิจัย

ตัวแปรมีทั้งหมด 9 แอททริบิวต์ แบ่งเป็นตัวแปรต้น 8 แอททริบิวต์ และตัวแปรตาม 1 แอททริบิวต์ตัวแปรต้นได้แก่ปัจจัยที่ใช้ในการจำแนกผู้ป่วยโรคเบาหวานมีจำนวน 8 แอตทริบิวต์ ได้แก่ การตั้งครรภ์ (Pregnancies), กลูโคส (Glucose), ความดันโลหิต (Blood pressure), ความหนาของชั้นผิวหนัง (Skin thickness), อินซูลิน (Insulin), ค่าดัชนีมวลกาย (Body mass index), ครอบครัวมีประวัติเป็นโรคเบาหวาน (Diabetes pedigree function), อายุ (Age) ตัวแปรตาม คือ ผลการจำแนก

ผู้ป่วยโรคเบาหวาน 0 แปลว่าไม่ใช่ผู้ป่วยโรคเบาหวาน 1 คือผู้ป่วยโรคเบาหวาน

1.2 ขอบเขตในการดำเนินการวิจัย

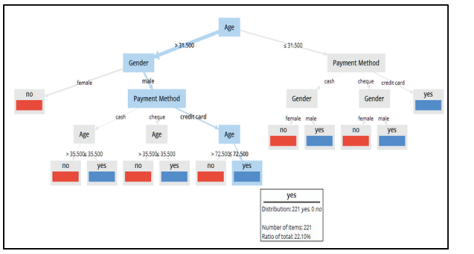
งานวิจัยนี้มุ่งเน้นการค้นหาเทคนิคด้านเหมืองข้อมูล เพื่อสร้างแบบจำลองเหมืองข้อมูลสำหรับการทำนายลักษณะจำแนกผู้ป่วยโรคเบาหวานโดยใช้อัลกอริทึมการเลือกลักษณะความสัมพันธ์ของข้อมูล เพื่อค้นหาอัลกอริทึมที่เหมาะสมที่สุดสำหรับฐานข้อมูลทางการแพทย์ โดยใช้อัลกอริทึมพื้นฐาน 4 อัลกอริทึม ซึ่งประกอบด้วย เนอีฟเบย์ (Naïve Bayes), ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine), เคเนียเรสเนเบอร์ (K-Nearest Neighbor), ต้นไม้ตัดสินใจ (Decision Tree) โดยทดสอบกับข้อมูลทางการแพทย์ผู้ป่วยโรคเบาหวาน

1.3 แนวคิดและทฤษฎีที่เกี่ยวข้อง

1.3.1 ต้นไม้ตัดสินใจ ต้นไม้จะประกอบด้วยโหนดแทนคุณลักษณะ และโหนดล่างสุดแทนหมวดหมู่การสร้างกิ่งสาขาจะพิจารณาจากค่าความจริงของคุณลักษณะ โดยค่าที่ใช้จะมาจากการคำนวณจากค่า Information Gain การสร้างต้นไม้ตัดสินใจ C4.5 ใช้ค่ามาตรฐานอัตราส่วนเกน (Gain Ratio) เพื่อเลือกคุณลักษณะที่จะใช้เป็นรากหรือโหนด ถ้าให้ชุดข้อมูล M ประกอบด้วยค่าที่เป็นไปได้ คือ $\{m_1, m_2, ..., m_n\}$ และให้ความน่าจะเป็นที่จะเกิดค่า m_1 มีค่าเท่ากับ P(mi) จะได้ว่าค่าเกนสารสนเทศ (Information Gain) ของ M เขียนแทนด้วย I(M) คำนวณได้ ดังสมการที่ (1) [2]

$$I(M)= \sum_{i=1}^{n} -P(m_i)\log_2 P(m_i) \qquad (1)$$

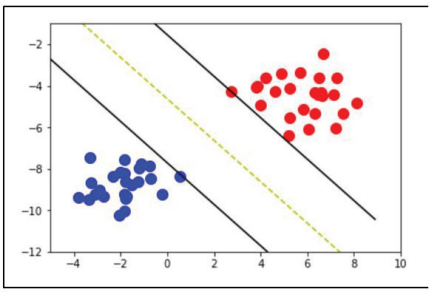
จากภาพโมเดลต้นไม้ตัดสินใจที่ได้จะสามารถระบุสาเหตุของปัญหาได้ โดยในงานวิจัยนี้ได้ทำการแบ่งคุณสมบัติออกเป็น 3 ประเภท คือ อายุ เพศ และระดับน้ำตาลในเลือด ดังแสดงในภาพที่ 1



ภาพที่ 1 ต้นไม้ตัดสินใจ

1.3.2 ซัพพอร์ทเวกเตอร์แมชชีน หลักการของวิธีการนี้ ใช้เพื่อหาระนาบการตัดสินใจในการแบ่งข้อมูลออกเป็นสองส่วน โดยใช้สมการเส้นตรงเพื่อแบ่งเขตข้อมูล 2 กลุ่มออกจากกัน ดังภาพที่ 2 โดยมีวัตถุประสงค์ที่จะพยายามที่จะทำการลดความผิดพลาดจากการทำนาย (Minimize error) พร้อมกับเพิ่มระยะแยกแยะให้มากที่สุด (Maximized Margin) ซึ่งต่างจากเทคนิคโดยทั่วไป เช่น โครงข่ายประสาทเทียม (Artificial Neural Network: ANN) ที่มุ่งเพียงทำให้ความผิดพลาดจากการทำนายให้ต่ำที่สุดเพียงอย่างเดียว โดยจะใช้ฟังก์ชั่นแม็ปข้อมูลจาก Input Space ไปยัง Feature Space และสร้างฟังก์ชั่นวัดความคล้ายที่เรียกว่าเคอร์เนลฟังก์ชั่น (Kernel Function) บน Feature Space เหมาะใช้สำหรับข้อมูลที่มีลักษณะมิติของข้อมูลที่มีปริมาณมาก โดยกำหนดให้ (xi,yi),...,(xn,yn) เป็นตัวอย่างที่ใช้สำหรับการสอน n คือ จำนวนข้อมูลตัวอย่าง m คือ จำนวนมิติข้อมูลเข้า และ y คือ ผลลัพธ์มีค่า + 1 หรือ -1 [2] ดังสมการ (xi,yi),...,(xn,yn) เมื่อ $x \in Rm$, $y \in \{+1,-1\}$

สำหรับปัญหาเชิงเส้น มิติข้อมูลขนาดสูงได้ถูกแบ่งเป็น 2 กลุ่ม โดยระนาบตัดสินใจ ซึ่งคำนวณได้ดังสมการ และเมื่อ w คือ ค่าน้ำหนักและ b คือ ค่า bias สมการ ใช้สำหรับจำแนกประเภทของข้อมูล (w*x)+b=0 โดย (w*x)+b>0 ถ้า yi = +1 และ (w*x)+b<0 ถ้า yi = -1 อย่างไรก็ตามซัพพอร์ทเวกเตอร์แมชชีนมีเคอร์เนลฟังก์ชั่น (Kernel Function) ที่ผู้สามารถประยุกต์ใช้ในการแก้ปัญหาได้หลายวิธีโดยผู้วิจัยต้องเลือกเคอร์เนลให้เหมาะสมกับลักษณะข้อมูล [9] [14]



ภาพที่ 2 ซัพพอร์ทเวกเตอร์แมชชีน [15]

1.3.3 เนอีฟเบย์ การเรียนรู้แบบเบย์ เป็นเทคนิคที่ใช้ทฤษฎีความน่าจะเป็นตามกฎของเบย์ (Bayes' Theorem) [3] เพื่อหาว่าสมมติฐานใดน่าจะถูกต้องที่สุด โดยใช้ความรู้ก่อนหน้า (Prior Knowledge) ได้แก่ ความน่าจะเป็นก่อนหน้าสำหรับสมมติฐานหนึ่ง ๆ ร่วมกับข้อมูล เช่น ความน่าจะเป็นที่สังเกตได้สำหรับสมมติฐานหนึ่ง ๆ เพื่อหาสมมติฐานที่ดีที่สุด การเรียนรู้แบบเบย์อาศัยหลักการของการคำนวณความน่าจะเป็นของแต่ละสมมติฐาน ในที่นี้คือคลาสเป้าหมายหรือผลลัพธ์การทำนายโดยการเรียนรู้แบบเบย์เป็นการเรียนรู้เพิ่มได้ เนื่องจากตัวอย่างใหม่ที่ได้มาถูกนำมาปรับเปลี่ยนการแจกแจงซึ่งมีผลต่อการเพิ่มหรือลดความน่าจะเป็นทำให้มีการเรียนรู้ที่เปลี่ยนไป วิธีการนี้ตัวแบบจะถูกปรับเปลี่ยนไปตามตัวอย่างใหม่ที่ได้โดยผนวกกับความรู้เดิมที่มี ซึ่งการทำนายค่าคลาสเป้าหมายของตัวอย่างใช้ความน่าจะเป็นมากที่สุดของทุกสมมติฐานจากทฤษฎีของเบย์ เราสามารถคำนวณความน่าจะเป็นของสมมติฐานต่าง ๆ โดยใช้สมการ ดังสมการที่ (2)

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \qquad (2)$$

P(h) คือ ความน่าจะเป็นก่อนหน้าของสมมติฐาน h

P(D) คือ ความน่าจะเป็นก่อนหน้าของชุดข้อมูลตัวอย่าง D

P(h|D) คือ ความน่าจะเป็นของ h ขึ้นต่อ D

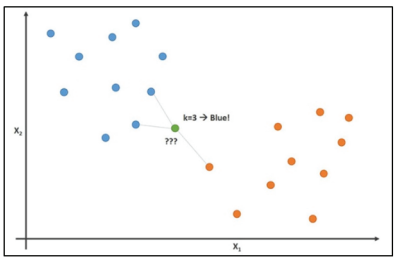
P(D|h) คือ ความน่าจะเป็นของ D ขึ้นต่อ h

ความน่าจะเป็นที่ B เกิดก่อน และ A เกิดตามมา

ความน่าจะเป็นที่ A และ B เกิดร่วมกัน [12] โดยใช้สมการ เนอีฟเบย์ ดังสมการที่ (3)

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad (3)$$

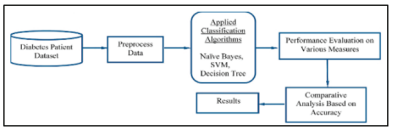
1.3.4 เคเนียเรสเนเบอร์ หลักการของวิธีการนี้จะจำแนกประเภทข้อมูลโดยขึ้นกับข้อมูลที่มีคุณสมบัติใกล้เคียงที่สุด K ตัวจากข้อมูลบนชุดข้อมูลตัวอย่าง ทำงานโดยขึ้นกับระยะทางน้อยสุดจากสมาชิกใหม่ หรือข้อมูลที่ป้อนถาม (input query instance) กับข้อมูลตัวอย่างฝึกฝน จะคำนวณหาเพื่อนบ้านที่ใกล้ที่สุด K ตัว หลังจากนั้นเราจะรวบรวมสมาชิกที่ใกล้เคียงที่สุด K ตัวแล้วเลือกคลาสที่สมาชิกส่วนใหญ่ที่ในกลุ่ม K ดังกล่าวสังกัดอยู่มากที่สุดให้กับสมาชิกใหม่ ข้อมูลการจำแนกโดยใช้ข้อมูลข้างเคียง K ตัว ประกอบด้วยแอททริบิวต์หลายตัวแปร Xi ซึ่งจะนำมาใช้ในการแบ่งกลุ่ม Yi โดยระบุค่าตัวเลขจำนวนเต็มบวกให้กับ K ซึ่งค่านี้จะเป็นตัวบอกจำนวนของกรณี (case) ที่จะต้องค้นหาในการทำนายกรณีใหม่ อัลกอริทึมแบบ K-NN ได้แก่ 1-NN , 2-NN , 3-NN , K-NN ตัวอย่าง 2-KNN หมายถึง อัลกอริทึมนี้จะค้นหา 2 กรณีที่มีลักษณะใกล้เคียงกับกรณีใหม่ (2 Nearest Cases) การนำระยะทางที่หาได้จากสมาชิกในข้อมูลตัวอย่างฝึกฝน มาเรียงลำดับจากน้อยไปหามากแล้วเลือกสมาชิกที่มีระยะทาง (Distance) ใกล้เคียงที่สุดออกมา K ตัวโดยใช้การวัดระยะทางแบบ Euclidean distance มีหลักการคือ การวัดระยะทางระหว่างสองวัตถุ ถ้าวัตถุห่างกันมากแสดงว่าวัตถุนั้นมีความคล้ายกันน้อย ถ้ามีค่าน้อยก็แสดงว่ามีความคล้ายคลึงกันมาก ดังภาพที่3 [4] [14]



ภาพที่ 3 เคเนียเรสเนเบอร์ [13]

1.3.5 การลดมิติข้อมูล (Feature Selection) เนื่องจากกลุ่มตัวอย่างของข้อมูลการแพทย์มีแนวโน้มที่จะเพิ่มปริมาณสูงขึ้นทุกวัน ทำให้ข้อมูลการแพทย์บางประเภทมีจำนวนคุณลักษณะมากขึ้น ซึ่งจำนวนคุณลักษณะมีผลต่อประสิทธิภาพของการจำแนกประเภทของโรคเนื่องจากอัลกอริทึมที่ใช้ในการเรียนรู้เพื่อสร้างตัวแบบจำลองการจำแนกผู้ป่วยโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูล โดยทั่วไปไม่สามารถรองรับการทำงานกับจำนวนคุณลักษณะของข้อมูลการแพทย์ที่สูงมากได้ดี การลดขนาดข้อมูลจึงเป็นขั้นตอนหนึ่งที่จะต้องทำก่อนการเรียนรู้ด้วยเครื่องจักรการเรียนรู้ (Machine learning) แต่การลดขนาดของคุณลักษณะข้อมูลการแพทย์ ต้องพิจารณาด้วยความระมัดระวัง เนื่องจากมีความเสี่ยงในการที่จะกำจัดคุณลักษณะที่สำคัญต่อการจำแนกประเภทของโรคออกไป จากการศึกษาพบว่าวิธีการลดคุณลักษณะมีหลากหลายวิธี ซึ่งอยู่บนพื้นฐานของการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรกับกลุ่มเป้าหมาย (Class) รวมถึงการสร้างคุณลักษณะใหม่จากคุณลักษณะเดิมอาจจะนำคุณลักษณะพื้นฐานเหล่านี้มารวมกันเพื่อให้เป็นคุณลักษณะใหม่

การลดมิติข้อมูลโดยใช้ความสัมพันธ์ของคุณลักษณะ (Correlation Based Feature Selection: CFS) อัลกอริทึมนี้มีหลักการกรองที่ง่าย ๆ โดย CFS จะจัดอันดับกลุ่มย่อยของมิติข้อมูล ซึ่งจะมีขั้นตอนการสร้างตัวจำแนกข้อมูล ดังภาพที่ 4 ตามความสัมพันธ์ที่อยู่บนพื้นฐานของฟังก์ชันการประมาณแบบ heuristic ซึ่งกลุ่มย่อยของมิติข้อมูล จะมีความสัมพันธ์กันสูงกับคลาส และไม่มีความสัมพันธ์กับคลาสอื่น ๆ สำหรับมิติข้อมูลที่ไม่เกี่ยวข้องอาจจะถูกละทิ้งเพราะมิติข้อมูลเหล่านี้อาจจะมีความสัมพันธ์ต่ำกับคลาส มิติข้อมูลที่ซ้ำซ้อนอาจจะถูกขจัดออกไปจากกลุ่มมิติข้อมูลที่มีความสัมพันธ์สูงสมการประเมินกลุ่มย่อยของมิติข้อมูลแบบ CFS [5]



ภาพที่ 4 ขั้นตอนการสร้างตัวจำแนกข้อมูล

2. วัตถุประสงค์ของการวิจัย

1. เพื่อสร้างแบบจำลองเหมืองข้อมูลสำหรับการทำนายลักษณะจำแนกผู้ป่วยโรคเบาหวานโดยใช้อัลกอริทึมการเลือกลักษณะความสัมพันธ์ของข้อมูลที่สามารถนำมาสนับสนุนการประยุกต์ใช้งานทางการแพทย์

2. เพื่อทดสอบและเปรียบเทียบประสิทธิภาพของแบบจำลองเหมืองข้อมูลสำหรับการทำนายลักษณะจำแนกผู้ป่วยโรคเบาหวานโดยใช้อัลกอริทึมการเลือกลักษณะความสัมพันธ์ของข้อมูล

3. การดำเนินการวิจัย

ในงานวิจัยนี้เริ่มต้นที่การนำข้อมูลมาทำการนอร์มัลไลซ์ (Normalization) โดยปรับค่าข้อมูลให้อยู่ในรูปแบบเดียวกันจากนั้นได้นำเสนอวิธีการสร้างตัวจำแนกข้อมูลที่หลากหลายโดยใช้ วิธีการสุ่มกลุ่มตัวอย่าง เพื่อสร้างความแตกต่างบนกลุ่มข้อมูลสำหรับการฝึกฝน (Training Sets) ซึ่งประกอบด้วยกลุ่มตัวอย่างจำนวน 768 คน ตัวจำแนกข้อมูลที่ได้กำหนดขึ้นสำหรับงานวิจัยนี้ได้ใช้การวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรมีรายละเอียด ดังภาพที่ 5 และใช้ตัวจำแนกทั้งสิ้น 4 ตัวประกอบ คือ Naïve Bayes, Support Vector Machine, Decision Tree และ K-Nearest Neighbor โดยใช้โปรแกรม RapidMiner มีลักษณะการทำงานของโปรแกรม ดังภาพที่ 6

โดยใช้ข้อมูลจาก UCI ซึ่งเป็นข้อมูลเปิดสำหรับนำไปทดสอบประสิทธิภาพอัลกอริทึมทางด้านเหมืองข้อมูลได้แก่ ฐานข้อมูล Diabetes Data [6]

การประเมินผลใช้การประเมินผลจาก

ประสิทธิภาพการจำแนกข้อมูล โดยวัดค่าความถูกต้องแม่นยำ (Accuracy) ในการจำแนกข้อมูลในกลุ่มข้อมูลทดสอบ ดังสมการที่ (4) [7]

$$\text{Accuracy} = \left(\frac{TP+TN}{TP+FN+TN+FP}\right) 100 \qquad (4)$$

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

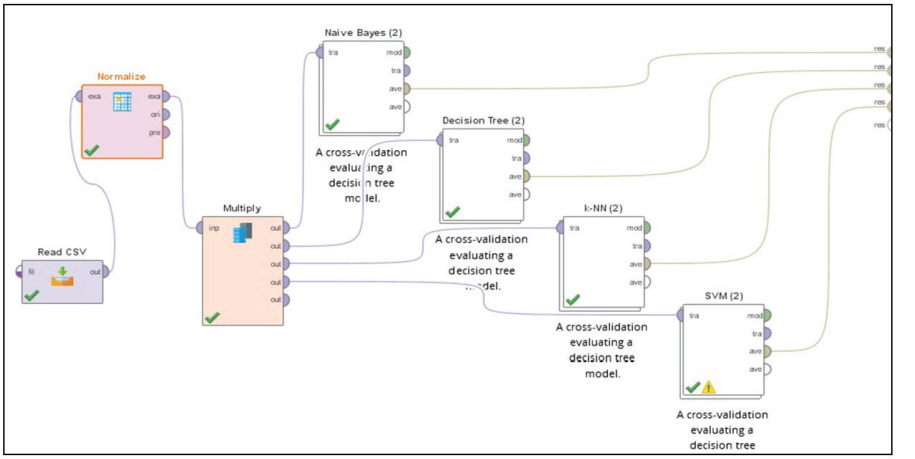
4. ผลการวิจัยและอภิปรายผลการวิจัย

กราฟที่ใช้แสดงการกระจายของตัวแปรแต่ละตัว ดังภาพที่ 7 พบว่าตัวแปร Pregnancies, Insulin, Diabetes Pedigree Function, Age และ Skin Thickness จะลาดลงไปทาง

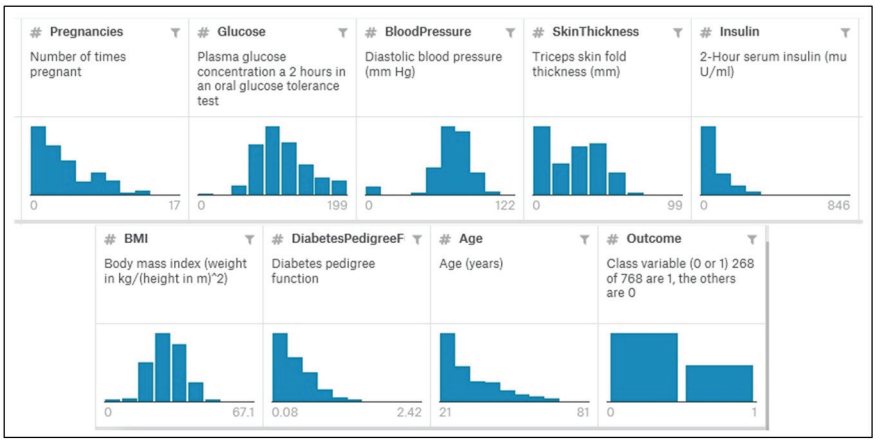
ค่าสูงมากกว่าปกติ ในลักษณะเบ้ขวา แต่ตัวแปร Glucose, BloodPressure, และ BMI ลักษณะการกระจายตัวเป็นไปตามปกติ ค่าเฉลี่ยส่วนใหญ่จะอยู่ตรงกลาง อีกทั้งสามารถแสดงผลการทดลองการจำแนกผู้ป่วยโรคเบาหวานโดยใช้อัลกอริทึม คือ Naïve Bayes, Support Vector Machine, Decision Tree และ K-Nearest Neighbor ซึ่งได้มีการเลือกลักษณะความสัมพันธ์ของข้อมูล คือ Precision, Recall, F-Measure และ Accuracy โดยแสดงผลเป็นค่าร้อยละของความสัมพันธ์ ดังภาพที่ 8

Attributes	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
Pregnancies	1	0.129	0.141	-0.082	-0.074	0.018	-0.034	0.544
Glucose	0.129	1	0.153	0.057	0.331	0.221	0.137	0.264
BloodPressure	0.141	0.153	1	0.207	0.089	0.282	0.041	0.240
SkinThickness	-0.082	0.057	0.207	1	0.437	0.393	0.184	-0.114
Insulin	-0.074	0.331	0.089	0.437	1	0.198	0.185	-0.042
BMI	0.018	0.221	0.282	0.393	0.198	1	0.141	0.036
DiabetesPedigreeFunction	-0.034	0.137	0.041	0.184	0.185	0.141	1	0.034
Age	0.544	0.264	0.240	-0.114	-0.042	0.036	0.034	1

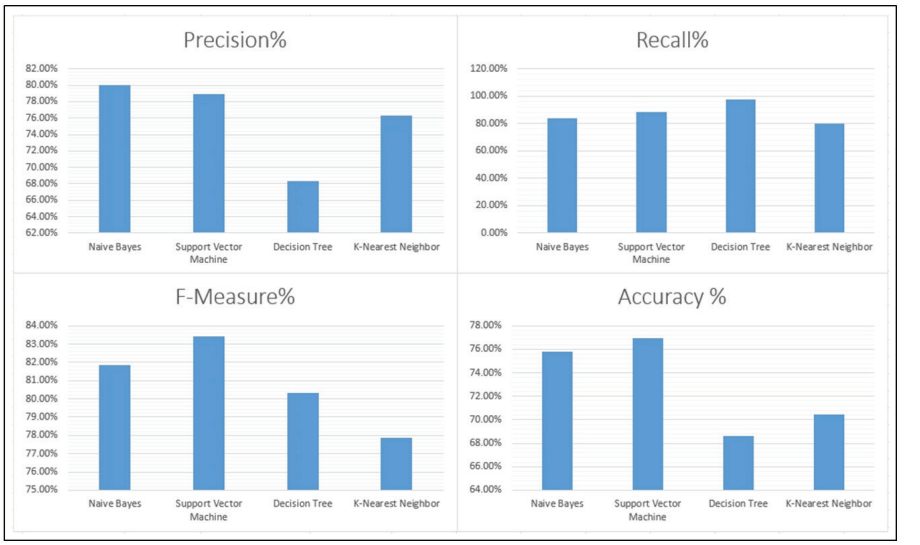
ภาพที่ 5 ความสัมพันธ์ของตัวแปร (Correlation Based Feature Selection)



ภาพที่ 6 ขั้นตอนการทำงานของโปรแกรม RapidMiner



ภาพที่ 7 ลักษณะการกระจายตัวของกลุ่มตัวอย่างในแต่ละตัวแปร



ภาพที่ 8 ผลการทดลองการจำแนกผู้ป่วยโรคเบาหวานโดยใช้อัลกอริทึมการเลือกลักษณะความสัมพันธ์ของข้อมูล

จากตารางผลการทดลองของต้นไม้ตัดสินใจจากข้อมูลชุดทดสอบของผู้ป่วยโรคเบาหวานจำนวน 768 คน พบว่าการทำนายลักษณะจำแนกผู้ป่วยโรคเบาหวานโดยใช้อัลกอริทึมต้นไม้ตัดสินใจนั้น ได้ค่า True Postive = 39, True Negative = 488, False Positive =12, False Negative = 229 โดยมีผลการทำนายค่าความถูกต้อง (Accuracy) เมื่อเทียบกับชุดข้อมูลของโมเดลต้นแบบ โดยเฉลี่ยร้อยละ 68.62% ให้ค่าความแม่นยำ (Precision) 68.27% ให้ค่าความระลึก (Recall) 97.60% ค่า F-Measure 80.34% ดังตารางที่ 1

ตารางที่ 1 ผลการทดลองของต้นไม้ตัดสินใจ

Accuracy: 68.62% +/- 3.72% (micro average: 66.62%)			
	True positive	True negative	Class precision
Pred. positive	39	12	76.47%
Pred. negative	229	488	68.06%
Class recall	14.55%	97.60%	

จากตารางผลการทดลองของซัพพอร์ทเวกเตอร์แมชชีน จากข้อมูลชุดทดสอบของผู้ป่วยโรคเบาหวานจำนวน 768 คน พบว่าการทำนายลักษณะจำแนกผู้ป่วยโรคเบาหวานโดยใช้อัลกอริทึมต้นไม้ตัดสินใจนั้น ได้ค่า True Postive = 149, True Negative = 442, False Positive = 58, False Negative = 119 โดยมีผลการทำนายค่าความถูกต้อง เมื่อเทียบกับชุดข้อมูลของโมเดลต้นแบบ โดยเฉลี่ยร้อยละ 76.95% ให้ค่าความแม่นยำ 78.96% ให้ค่าความระลึก 88.40% ค่า F-Measure 83.41% ดังตารางที่ 2

ตารางที่ 2 ผลการทดลองของซัพพอร์ทเวกเตอร์แมชชีน

Accuracy: 76.95% +/- 5.39% (micro average: 76.95%)	True positive	True negative	Class precision
Pred. positive	149	58	71.98%
Pred. negative	119	442	78.79%
Class recall	55.60%	88.40%	

จากตารางผลการทดลองของเนอีฟเบย์ จากข้อมูลชุดทดสอบของผู้ป่วยโรคเบาหวานจำนวน 768 คน พบว่าการทำนายลักษณะจำแนกผู้ป่วยโรคเบาหวานโดยใช้อัลกอริทึมต้นไม้ตัดสินใจนั้น ได้ค่า True Postive = 163, True Negative = 419, False Positive = 81, False Negative = 105 โดยมีผลการทำนายค่าความถูกต้อง เมื่อเทียบกับชุดข้อมูลของโมเดลต้นแบบ โดยเฉลี่ยร้อยละ 75.79% ให้ค่าความแม่นยำ 79.99% ให้ค่าความระลึก 83.80% ค่า F-Measure 81.85% ดังตารางที่ 3

ตารางที่ 3 ผลการทดลองของเนอีฟเบย์

Accuracy: 75.79% +/- 6.04% (micro average: 75.78%)	True positive	True negative	Class precision
Pred. positive	163	81	66.80%
Pred. negative	105	419	79.96%
Class recall	60.82%	83.80%	

จากตารางผลการทดลองของเคเนียเรสเนเบอร์ จากข้อมูลชุดทดสอบของผู้ป่วยโรคเบาหวานจำนวน 768 คน พบว่าการทำนายลักษณะจำแนกผู้ป่วยโรคเบาหวานโดยใช้อัลกอริทึมต้นไม้ตัดสินใจนั้น ได้ค่า True Postive = 143, True Negative = 398, False Positive = 102, False Negative = 125 โดยมีผลการทำนายค่าความถูกต้อง เมื่อเทียบกับชุดข้อมูลของโมเดลต้นแบบ โดยเฉลี่ยร้อยละ 70.45% ให้ค่าความแม่นยำ 76.25% ให้ค่าความระลึก 79.60% ค่า F-Measure 77.89% ดังตารางที่ 4

ตารางที่ 4 ผลการทดลองของเคเนียเรสเนเบอร์

Accuracy: 70.45% +/- 4.73% (micro average: 70.44%)	True positive	True negative	Class precision
Pred. positive	143	102	58.37%
Pred. negative	125	398	76.10%
Class recall	53.36%	79.60%	

งานวิจัยนี้เป็นการสร้างแบบจำลองเหมืองข้อมูลสำหรับการทำนายลักษณะจำแนกผู้ป่วยโรคเบาหวาน โดยใช้อัลกอริทึมการเลือกลักษณะความสัมพันธ์ของข้อมูล โดยมีวัตถุประสงค์เพื่อสร้างและเปรียบเทียบประสิทธิภาพของแบบจำลองเหมืองข้อมูลสำหรับการทำนายลักษณะจำแนกผู้ป่วยโรคเบาหวาน ผลการทดลองสำหรับการทำนายลักษณะจำแนกของผู้ป่วยโรคเบาหวาน โดยใช้แบบจำลองเหมืองข้อมูลที่สร้างโดยใช้เทคนิคซัพพอร์ทเวกเตอร์แมชชีน มีประสิทธิภาพการทำนายสูงสุดคิดเป็น 76.95% ซึ่งสามารถนำผลที่ได้ไปประยุกต์ใช้ในการสร้างระบบสนับสนุนการตัดสินใจในส่วนของคัดกรองโรคและแนวทางการรักษาของแพทย์และผู้ป่วยได้ต่อไป

5. เอกสารอ้างอิง

[1] อดุลย์ ยิ้มงาม (2560). *การทำเหมืองข้อมูล Data Mining.* [ออนไลน์], สืบค้นจาก http://compcenter.bu.ac.th/index.php?option=com_content&task=view&id=75&Itemid=172. (25 มีนาคม 2560).

[2] พรพล ธรรมรงค์รัตน์, ลัดดา ปรีชาวีรกุล, และ วิภาดา เวทย์ประสิทธิ์. (2551). การจำแนกประเภทเว็บเพจโดยใช้ค่าความถี่เอกสารและซัพพอร์ตเวกเตอร์แมชชีน. ใน *การประชุมวิชาการวิทยาการคอมพิวเตอร์และวิศวกรรมคอมพิวเตอร์แห่งชาติ ครั้งที่ 12.* ชลบุรี.

[3] Bayes, T., & Price, R. (1763). An Essay towards solving a problem in the Doctrine of chance. *Philosophical Transactions of the Royal Society of London*, 53, 370-418.

[4] บุญเสริม กิจศิริกุล. (2546). *อัลกอริทึมการทำเหมืองข้อมูล* (รายงานวิจัยฉบับสมบูรณ์). กรุงเทพฯ: จุฬาลงกรณ์มหาวิทยาลัย.

[5] นรินทร์ พนาวาส, และ นิเวศ จิระวิชิตชัย. (2553). การจำแนกมะเร็งเม็ดเลือดขาวโดยใช้เทคนิคการลดมิติข้อมูลด้วย Chi-square (หน้า 7-12). ใน *การประชุมวิชาการระดับประเทศด้านเทคโนโลยีสารสนเทศ ครั้งที่ 3 (NCIT10).*

[6] UCI Machine Learning Repository. (2008). *Pima Indians Diabetes Database*. [Online], Retrieved from https://www.kaggle.com/uciml/pima-indians-diabetes-database.

[7] Chirawichitchai, N. (2012). Data classification Using Weka software. *Journal of Engineering Siam University.*

[8] วิธวินท์ แสงมณี. (2557). *การสร้างแบบจำลองทำนายโอกาสการกลับมารักษาตัวซ้ำของผู้ป่วยโรคเบาหวาน.* ประจวบคีรีขันธ์: มหาวิทยาลัยเทคโนโลยีราชมงคลรัตนโกสินทร์ วิทยาเขตวังไกลกังวล.

[9] Han, J., & Kamber, M. (2006). Data *Mining: Concepts and techniques* (2nd ed.). San Francisco: Morgan Kaufmann.

[10] RapidMiner (2019). *What's New in RapidMiner Studio 7.5?*. [Online], Retrieved from https://docs.rapidminer.com/7.6/studio/releases/7.5/. (2019, June 30)

[11] เอกสิทธิ์ พัชรวงศ์ศักดา (2557). *เทคนิค Support Vector Machie (SVM) และซอฟต์แวร์ HR-SVM.* [ออนไลน์], สืบค้นจาก http://dataminingtrend.com/2014/support-vector-machine-svm/. (13 มีนาคม 2557).

[12] เอกสิทธิ์ พัชรวงศ์ศักดา (2557). *โมเดล Naïve Bayes และการแปลความหมาย*. [ออนไลน์]. สืบค้นจากhttp:// dataminingtrend.com/2014/naive-bayes. (17 มีนาคม 2557).

[13] Kishor, N. (2018). *K-Nearest Neighbors – the Laziest Machine Learning Technique.* [Online], Retrieved from https://www.houseofbots.com/news-detail/2542-4-k-nearest-neighbors-the-laziest-machine-learning-technique. (2018, April 2).

[14] Chirawichitchai, N. (2013). Automatic Thai Document Classification Model. *Journal of Industrial Technology King Mongkut's University of Technology*, 9, 141–149.

[15] Mr. P L. (2561). *SVM อดีตเคยหวานปัจจุบันแอบเซง : Machine Learning 101*. [ออนไลน์], สืบค้นจาก https://medium.com/mmp-li/svm-อดีตเคยหวานปัจจุบันแอบเซง-machine-learning-101-6008753c780c. (15 พฤศจิกายน 2561).